SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 1-33, January 1990

INVITED EXPOSITORY ARTICLE

This paper is another in the continuing series of expository papers that were invited by the editors. These papers undergo the same refereeing procedure as do research papers submitted directly by the authors, although the refereeing guidelines are modified to suit the largely expository nature of the paper. Due to the rapid recent technical development of a number of areas in control and optimization, many of the seminal papers are quite specialized and are readily accessible to a limited group of experts only. Moreover, the original motivations and practical importance of the ideas are sometimes difficult to find in the mathematical development. The purpose of these papers is to bring the ideas, techniques, and applications of a few selected areas to the attention of a wider audience, so that their basic importance can be more easily and widely appreciated.

CONTROLLABILITY OF NONLINEAR DISCRETE-TIME SYSTEMS: A LIE-ALGEBRAIC APPROACH*

BRONISLAW JAKUBCZYK† AND EDUARDO D. SONTAG‡

Abstract. This paper presents a geometric study of controllability for discrete-time nonlinear systems. Various accessibility properties are characterized in terms of Lie algebras of vector fields. Some of the results obtained are parallel to analogous ones in continuous-time, but in many respects the theory is substantially different and many new phenomena appear.

Key words. controllability, Lie algebras of vector fields, nonlinear systems, discrete time

AMS(MOS) subject classifications. 93C10, 93C55, 93B05

1. Introduction. This paper deals with questions of controllability for discrete-time nonlinear systems

$$x(t+1)=f(x(t), u(t)) \tag{1}$$

for which the control variables u and state variables x take continuous values. Systems of the type (1) but with discrete-valued states and controls have long been studied in automata and sequential machine theory, but the continuous case has only recently become the subject of serious investigation as far as controllability properties are concerned. Our objective here is to survey a number of known results and to present new characterizations involving geometric ideas.

The study of controllability questions for the better known continuous-time analogue of (1), the differential equation

$$\dot{x}(t)=\phi(x(t), u(t)), \tag{2}$$

has been the subject of a concentrated research effort, as documented, for instance, in the survey papers [2] and [7], the text [8], and the exposition [35]. It is known, for instance, that the set *accessible* from any given state x^0, that is to say, the set of points reachable from x^0, contains a smooth submanifold of the state space and is in turn contained in a submanifold of the same dimension. Thus, for instance, the cusp in Fig. 1 cannot be an accessible set for any system of the type (2). More interestingly perhaps, this dimension can be computed from the rank of certain matrices formed

* Received by the editors January 11, 1988; accepted for publication (in revised form) March 14, 1989.
† Institute of Mathematics, Polish Academy of Sciences, Sniadeckich 8, 00-950 Warsaw, Poland. This work was done while the author was a Visiting Professor at Rutgers, The State University of New Jersey.
‡ Rutgers Center for Systems and Control, Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903. E-mail address: sontag@math.nutgens.edu. This research was supported in part by United States Air Force grant 85-0247 and National Science Foundation grant DMS-8803396.

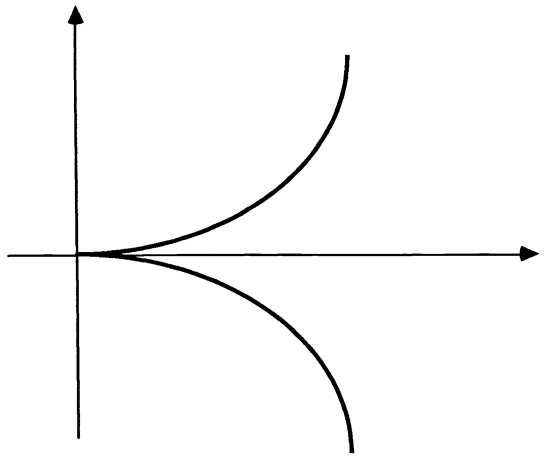


FIG. 1. *Impossible reachable set.*

by taking iterated Jacobians of the various vector fields $\phi(\cdot, u)$ evaluated at the state x^0. These Lie-theoretic characterizations are "direct" in that they do not involve integration of the differential equation, and they are closely related to more classical geometric material related to Frobenious' theorem.

(Certain technical hypotheses are of course required for the validity of the above and other assertions that we will make here; for purposes of providing an informal introduction we shall not make them precise yet; however, as a general rule, real-analyticity of f and ϕ and the assumption that states and controls take values in Euclidean space $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, are more than sufficient.)

Discrete control systems (1) are of interest for various reasons. Of course in many areas difference equation models are more natural than differential equations, but our interest has been motivated more by the problem of modeling physical systems under digital control via *sampling*. Recall that sampling is the process under which the state of a continuous time system is measured at discrete instants, and control actions are taken also at discrete instants. Figure 2 illustrates a typical approach to computer control. A discrete-time algorithm observes the state (or more generally, the outputs)

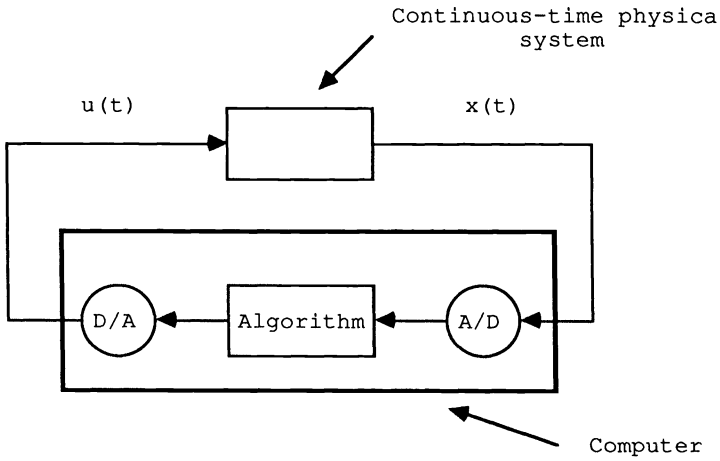


FIG. 2. *Digital control configuration.*

of a physical system, through an analogue-to-digital converter. Typically this observation is made at periodic time instants $\delta, 2\delta, \cdots$. On the basis of this observation the controller decides upon a control value u to be applied during the next period of length δ. This value is converted to analogue form and is held constant during that next period. So the controls applied to the physical system are restricted to be δ-sampled controls, constant on intervals $[k\delta, (k+1)\delta]$ (Fig. 3). The main point here is that, as far as the control algorithm is concerned, the physical system is a discrete-time system described by an equation of type (1), where $f(x, u)$ is the solution of the differential equation (2) at the end of an interval of length δ assuming that the initial state was x and control was held constantly equal to u.

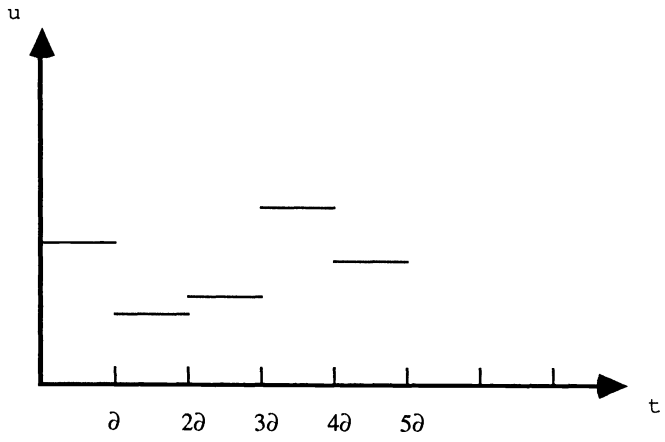


FIG. 3. *∂-sampled control.*

This description of sampling is oversimplified in many respects. For instance, analogue/digital conversion involves a quantization of the values of x into a discrete number of steps. Constant controls values may be smoothed out by a filter before being applied to the system. Multirate strategies, in which the sampling period is varied in a fixed set, may also be used. And the time involved in the algorithm actually computing the value of the control is sometimes nontrivial and must be included in the model as well. But even without these complications, the study of discrete-time control systems appears naturally.

Another area in which results from discrete-time nonlinear control theory are of importance is in the study of Markovian systems (1). There, the variables $u(t)$ are random, and together with the transitions f they characterize the probabilistic behavior of the process $x(\cdot)$. Accessibility conditions play a central role in establishing the existence and smoothness properties of equilibrium distributions; see for instance [15] and [16].

Yet another source of discrete-time control systems, related to but different from sampling, arises when numerically approximating the solution of a system (2). For instance, a Euler approximation with stepsize h gives the recursion

$$x(t+1) = x(t) + h\phi(x(t), u(t)).$$

These motivations notwithstanding, discrete-time systems have been studied much less than their continuous counterparts, and it has long been felt that their properties

may diverge considerably from those of the latter. Regarding control and observation problems, the paper [26] and the monograph [27] considered various aspects of discrete-time systems defined by polynomial evolution equations. However, the general theory remained, until recently, much weaker than that possible in the more classical continuous time case, for which a large body of knowledge, as described above, is now available.

One of the main difficulties in the general discrete-time case is due to the possible noninvertibility of the one-step transition maps

$$x \mapsto f(x, u),$$

which means that semigroups tend to appear where groups would appear in the continuous case, so less algebraic structure is available. Accessible sets with singularities such as the curve in Fig. 1 can then easily appear.

An important observation, however, is that—due to the time-reversibility of finite-dimensional differential equations—for those discrete-time systems that arise through sampling these transition maps, obtained by integrating (2) over an interval of length δ with control $\equiv u$, *are* invertible. More precisely, each of these maps is a diffeomorphism (possibly not everywhere defined) of the state space. This is analogous to the situation in classical dynamical system theory, where one studies time-one diffeomorphisms and Poincaré maps associated to differential equations. Invertible discrete-time systems are often also obtained in numerical schemes for discretizing continuous-time models, if mesh sizes are chosen small enough.

In this paper we shall restrict our attention to *invertible* systems, for which the maps $f(\cdot, u)$ are assumed to be diffeomorphisms. For such systems we derive several characterizations of accessibility and we study the geometric structure of accessible sets. As an example, we provide a theorem that shows that, at least from equilibrium states, a picture such as that in Fig. 1 can never hold for these sets. (Precise statements of results are given later.) As with continuous-time systems, we also give Lie-theoretic characterizations of accessibility. These characterizations have the advantage that they do not require the computation of arbitrary iterates of the transition map, save for those iterates corresponding to just one value of the control value set.

The basic fact that underlies our approach is that one has an analogue for difference equations of the infinitesimal information obtained in the continuous-time case by taking derivatives with respect to time. One uses here derivations *with respect to control values.* This idea can be traced back to the paper [9], the first to deal in detail with general invertible discrete nonlinear control systems, although in the context of realization theory rather than controllability problems. For the latter, and for the source of the closest related material to that presented here, the credit goes to Fliess and Normand-Cyrot ([3], [25]), who originally proposed the definition in this manner of Lie algebras associated to discrete-time systems. This is analogous to associating a Lie algebra action to any given Lie group action. Other work along those lines was carried out in [11], [32], [17], [29], and related papers. A particularly important line of work is that pursued in [18], [20], [22], as well as by other authors (see, e.g., [5]), who have shown how to frame a large number of problems of control design (decoupling, noninteracting control, immersion, and so forth) in this geometric formalism; we shall not deal with such questions in this paper, however. For other recent references on geometric discrete-time control, see, for instance, the following papers as well as references given there: [1], [6], [10], [12], [14], [19], [24], [28].

We close this introduction with the precise statement of a simplified version of one of our main results to illustrate the nature of our contribution. Assume that the

system (1) is analytic, in the sense that f is analytic, and invertible, meaning that each of the maps

$$f_u = f(\cdot, u): \mathbb{R}^n \to \mathbb{R}^n$$

is a global diffeomorphism of $\mathbb{R}^n$ for each control value u; for simplicity assume further that the control values are arbitrary real numbers, $u \in \mathbb{U} := \mathbb{R}$.

Denote by f_0^k the kth power of f_0 with respect to composition, and define the following vector fields depending on u:

$$X_u^+(x) = \frac{\partial}{\partial v}\bigg|_{v=0} f_u^{-1} \circ f_{u+v}(x),$$

$$X_u^-(x) = \frac{\partial}{\partial v}\bigg|_{v=0} f_u \circ f_{u+v}^{-1}(x),$$

and more generally for each integer k and for $\sigma = \pm$, $f_u^+ = f_u$, $f_u^- = f_u^{-1}$,

$$(\mathrm{Ad}_0^k X_u^\sigma)(x) = \frac{\partial}{\partial v}\bigg|_{v=0} f_0^{-k} \circ f_u^{-\sigma} \circ f_{u+v}^{\sigma} \circ f_0^k(x),$$

where $-\sigma = -$, $+$ if $\sigma = +$, $-$, respectively. These vector fields were introduced in [11], [17], [20], and [21].

In analogy with standard continuous time notions of accessibility, we call the system (1) *forward accessible* from the state $x^0 \in \mathbb{R}^n$ if its attainable set from x^0 has a nonempty interior. Similarly, we say that (1) is *backward accessible* from x^0 it its backward attainable set from x^0, the set of points controllable to x^0, has a nonempty interior. Finally, we say that the system is *forward-backward accessible* or *transitive* from x^0 if its orbit through this state (the smallest positive and negative-invariant set containing x^0) has a nonempty interior. The orbit turns out to be a submanifold, so forward-backward accessibility is equivalent to this orbit being an open subset of the state space.

By an equilibrium state x^0 we mean one that satisfies $f(x^0, 0) = 0$. Part (c) of the following theorem had already been stated in [11] (see also Theorem 7 in [20]) but parts (a) and (b) are totally new. The theorem is a specialization to analytic systems and equilibrium states of much more general results to be discussed later.

THEOREM 1. *The following statements hold for any analytic system* (1) *and equilibrium state* x^0:

(a) *System* (1) *is forward accessible from* x^0 *if and only if*

$$\dim \mathrm{Lie}\,\{\mathrm{Ad}_0^k X_u^+ | k \geqq 0,\ u \in \mathbb{U}\}(x^0) = n.$$

(b) *System* (1) *is backward accessible from* x^0 *if and only if*

$$\dim \mathrm{Lie}\,\{\mathrm{Ad}_0^k X_u^- | k \leqq 0,\ u \in \mathbb{U}\}(x^0) = n.$$

(c) *System* (1) *is forward-backward accessible from* x^0 *if and only if*

$$\dim \mathrm{Lie}\,\{\mathrm{Ad}_0^k X_u^\sigma | k \in \mathbb{Z},\ u \in \mathbb{U},\ \sigma = \pm\}(x^0) = n.$$

It is an easy corollary of this theorem that all three conditions (forward, backward, and forward-backward accessibility) coincide for analytic systems and equilibrium initial states. This gives a generalization of the well-known Chow Theorem in the continuous-time theory. More generally, the dimension of the corresponding (forward, etc.) accessible sets are given by the dimensions of the above subspaces, from which it follows that the (forward) accessible set is an open subset of a manifold (the orbit);

therefore, the cusp in Fig. 1 cannot be a forward accessible set. Later we give an example for which this cusp appears as the union of three orbits, corresponding to the origin and each of the two smooth branches.

Note that the conditions in Theorem 1 involve iterated compositions of transitions corresponding to only *one* control—arbitrarily taken as the zero control. The "naive" conditions that one can give based on the implicit function theorem for the above accessibility properties, reviewed below, would involve compositions of all transition mappings, as well as, for backward and forward-backward accessibility of their (possibly hard to compute) inverses. Moreover, in the particular case when the system has, for instance, the form

$$x(t+1) = x(t) + g(x(t), u(t))$$

with $g(x, 0) \equiv 0$, the "Ad's" become all the identity and no compositions at all need be computed.

In this paper, we present an exposition, including complete proofs, of the known transitivity (positive and negative-time accessibility) facts, as well as of new results for the substantially different (positive-time) forward accessibility problem. We also clarify the relationship between a large number of forward and/or backward controllability notions. Another topic studied is the role played by various continuous time systems derived mathematically from the original discrete time model, and we show how to view the more classical results for continuous-time systems as a particular case (essentially when "time" is thought of as a control) of our theory. Finally, we provide an application of our accessibility characterizations to the sampled control of continuous systems; the resulting explicit eigenvalue condition, which generalizes the classical (linear system) sampling theorem, illustrates the power of the techniques developed. An illustrative example is included towards the end of the paper, which ends with a brief description of the alternative approach due to Normand-Cyrot.

2. Basic definitions. We start by introducing basic notation and definitions. As stated previously, time takes integer values, $t \in \mathbb{Z}$. We introduce the following notations for the effect of shift operators:

$$x^+(t) = x(t+1) \quad \text{and} \quad x^-(t) = x(t-1).$$

In this way we can write equation (1) in the more compact form, with $f^+ = f$

$$x^+ = f^+(x, u), \qquad x(t) \in \mathbb{X},\ u(t) \in \mathbb{U}.$$

The state set $\mathbb{X}$ is a connected differentiable manifold of dimension n. To simplify the notation we first assume that the control is scalar, meaning that $\mathbb{U}$ is a subset of $\mathbb{R}$ contained in the closure of its interior,

$$\mathbb{U} \subset \operatorname{clos} \operatorname{int} \mathbb{U},$$

such that $0 \in \mathbb{U}$. Later we show how to generalize everything to the case where $\mathbb{U}$ is a subset of a more general manifold.

The system is *of class* C^k if the manifold $\mathbb{X}$ is of class C^k, Hausdorff, second countable, and the function $f: \mathbb{X} \times \mathbb{U} \to \mathbb{X}$ is of class C^k, meaning, to be precise, that there exists a C^k extension of f to an open neighborhood of $\mathbb{X} \times \mathbb{U}$ in $\mathbb{X} \times \mathbb{R}$. When $k = \infty$ we say simply *smooth*; for $k = \omega$, *analytic*.

Associated to each such system there is a family of maps

$$f_u = f(\cdot, u): \mathbb{X} \to \mathbb{X}, \qquad u \in \mathbb{U}.$$

DEFINITION 2.1. The system (1) is *invertible* if for each u in an open neighborhood of $\mathbb{U}$ the map f_u is a global diffeomorphism of $\mathbb{X}$.

Invertibility can be weakened in various ways. For instance, many results can be obtained under the assumption of *local invertibility at x*, meaning that for each $u \in \mathbb{U}$ f_u is a local diffeomorphism at x, i.e., rank $(\partial f_u/\partial x)(x) = n$, or the assumption that this holds for every state, *local invertibility* of the system. The paper [10] shows how a condition called *submersibility* is in fact enough to define many of the concepts that we use in this paper.

To any invertible system one can associate an *inverse* or *reversed-time* system with equations

$$x^- = f^-(x, u), \tag{3}$$

where $f^-(x, u) = f_u^{-1}(x)$. By the implicit mapping theorem, this is again of class C^k, and its inverse is the original system.

Unless otherwise stated, every system appearing in this paper will be assumed to be invertible. Furthermore, until § 6, *controls are scalar.*

The maps f_u and their inverses f_u^{-1} can be considered as "one step forward maps" (respectively, "one step backward maps"). If we apply a sequence of controls $u_1, \cdots, u_k$ then we obtain the composition of these maps denoted by

$$f_{u_k \cdots u_1} = f_{u_k} \circ \cdots \circ f_{u_1}. \tag{4}$$

Allowing backward as well as forward steps we obtain a larger family of maps

$$f_{u_k \cdots u_1}^{\varepsilon_k \cdots \varepsilon_1} = f_{u_k}^{\varepsilon_k} \circ \cdots \circ f_{u_1}^{\varepsilon_1}, \tag{5}$$

where each of $\varepsilon_1, \cdots, \varepsilon_k$ takes a value ± 1.

We shall denote by $A_k^+(x)$ the set of points attainable from x in k forward steps, and by $A^+(x)$ the set of points attainable from x in any nonnegative number of forward steps. Replacing forward steps by backward steps we obtain other sets, $A_k^-(x)$ and $A^-(x)$, which consist of points controllable to x in k steps, and controllable to x in any nonnegative number of steps, respectively. Finally, the set of points attainable from x in any number of positive and negative steps is called the *orbit* of x and is denoted by $A(x)$.

DEFINITION 2.2. The system (1) is *forward* (*backward*) *accessible from x* if its attainable set $A^+(x)$ (respectively, $A^-(x)$) has a nonempty interior. It is called *transitive from x* (*or forward-backward accessible from x*) if its orbit $A(x)$ has a nonempty interior (and so it is necessarily open).

Finally, the system is *forward* (*backward*) *accessible* if it is forward (backward) accessible from any $x \in \mathbb{X}$, and it is called *transitive* if it is transitive from any $x \in \mathbb{X}$.

Observe that there is a straightforward criterion for accessibility of the discrete time system, based on the rank of the following map. For each fixed state x and integer k define

$$\psi_{k,x}(\mathbf{u}) := f_{u_k \cdots u_1}(x),$$

where $\mathbf{u} = (u_1, \cdots, u_k)$ takes values in the kth Cartesian product $\mathbb{U}^k$. Notice that the attainable set $A_k^+(x)$ is by definition equal to the image of this map. The following proposition says that this set is of nonempty interior if and only if the linearization along some trajectory starting from x is controllable.

PROPOSITION 2.3. *Let* (1) *be smooth. For any fixed* x *and* k, *the interior of the attainable set* $A_k^+(x)$ *is nonempty if and only if*

$$\sup\left\{\operatorname{rank}\frac{\partial}{\partial \mathbf{u}}\psi_{k,x}(u), \mathbf{u}\in \mathbb{U}^k\right\} = n$$

and thus

$$\sup\left\{\operatorname{rank}\frac{\partial}{\partial \mathbf{u}}\psi_{k,x}(u), \mathbf{u}\in \mathbb{U}^k, k\geqq 1\right\} = n$$

is necessary and sufficient for forward accessibility of system (1) *from* x.

Proof. If there is a point $\mathbf{u}$ at which the rank of the map $\psi_{k,x}$ is equal to n, we may assume without loss of generality that $\mathbf{u}$ is in the interior of $\mathbb{U}$, because of the hypothesis that $\mathbb{U}\subset \operatorname{clos}\operatorname{int}\mathbb{U}$. It then follows from the implicit function theorem that the image of this map has a nonempty interior. Thus, the attainable set $A_k^+(x)$ has a nonempty interior. (Only that the system is of class C^1 is used for this implication.)

Conversely, if the rank of the map $\psi_{k,x}$ is less than n at each $u\in\mathbb{U}$, then every element of $A_k^+(x)$ is a critical value of $\psi_{k,x}$ as a map defined on an open subset of $\mathbb{R}^k$. It follows by Sard's theorem that the image of $\mathbb{U}$ under this map is of empty interior and is of measure zero under the measure induced by any Riemann metric on $\mathbb{X}$ (the Euclidean metric in $\mathbb{R}^n$). Therefore, the attainable set $A_k^+(x)$ must have an empty interior and it is even of measure zero.

The second statement follows from the first because a countable union of sets of measure zero again has measure zero. □

REMARK 2.4. Since the orbit $A(x)$ is the (countable) union of the images of the maps (5) we can use an analogous argument to give a criterion for transitivity from x, using the maps (5) rather than (4) to define a family of maps playing the role of the $\psi_{k,x}$'s.

The above proposition and remark might appear to give satisfactory criteria for forward accessibility and transitivity. Unfortunately, this is not the case. Although for simple systems they may be used to decide whether a given system is forward accessible or not, for more complicated sytems explicitly computing the functions $\psi_{k,x}$ may be highly nontrivial, since composition is hard to deal with computationally. As an example, consider for instance the problem of obtaining a general formula for the nth composition of the quadratic function $g(x)=ax^2+bx+c$ with itself or that of computing the function $\psi_{k,x}$ if $f(x,u)=g(x)+xu$. The problem becomes even more serious in the case of deciding the transitivity of the system, as this requires also finding the inverse maps f_u^{-1} needed for computing the composed maps (5). One approach here is to develop a calculus for these compositions, as in the work of Monaco and Normand-Cyrot; see the last section. But in any case, even for classes such as that of bilinear systems, Proposition 2.3 doesn't seem to provide much useful information regarding accessibility properties.

Also, from a purely theoretical point of view, Proposition 2.3 is of little interest. This is because it gives too limited an insight into the geometry of our systems and it provides an even more limited tool for their study. The maps appearing in the criteria do not have much algebraic and geometric structure.

The main aim of the next section is to introduce a sort of "infinitesimal description" of the discrete-time system. This is done by introducing certain vector fields associated to it. By doing so we immediately get a powerful tool and a rich algebraic and geometric structure based on the Lie product of vector fields. In particular, the accessibility properties of the system can be studied using natural Lie algebras of vector fields

associated to the system. The idea of introducing vector fields corresponding to infinitesimal perturbations of control values is a natural generalization of the concept of actions of Lie groups, and it was originally proposed in the context of nonlinear control in [3]. These vector fields also find natural applications in the study of controllability properties and the feedback linearizability of sampled systems ([29], [12]).

3. Vector fields associated to the system. We associate the following four families of vector fields to our discrete time system (1), one vector field for each $u \in \mathbb{U}$:

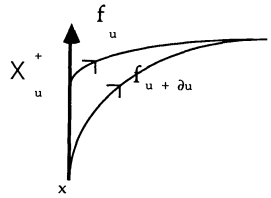
$$X_u^+(x) = \frac{\partial}{\partial v}\bigg|_{v=0} f_u^{-1} \circ f_{u+v}(x),$$

$$X_u^-(x) = \frac{\partial}{\partial v}\bigg|_{v=0} f_u \circ f_{u+v}^{-1}(x),$$

$$Y_u^+(x) = \frac{\partial}{\partial v}\bigg|_{v=0} f_{u+v}^{-1} \circ f_u(x),$$

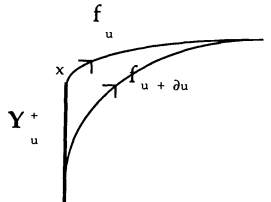
$$Y_u^-(x) = \frac{\partial}{\partial v}\bigg|_{v=0} f_{u+v} \circ f_u^{-1}(x).$$

The partial derivatives here are well defined in the interior of $\mathbb{U}$; therefore, they are also uniquely defined on the boundary of $\mathbb{U}$ because of continuity. The geometric



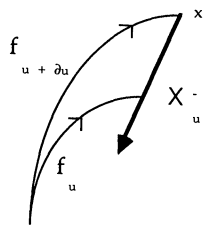
X_u^+

FIG. 4 (a)



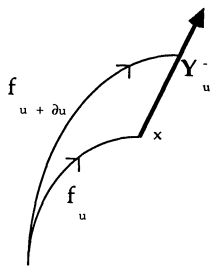
Y_u^+

FIG. 4 (c)



X_u^-

FIG. 4 (b)



Y_u^-

FIG. 4 (d)

meaning of these vector fields is illustrated by Fig. 4, and the interrelations between them are explained in the next proposition. These vector fields were also introduced in [17], [20], and [21], using somewhat different terminology. The last section will explain the relation between the different notations.

The special case in which the function f happens to correspond to the flow of a vector field Z, that is, $f(x, u) = \exp(uZ)$, will be important later when discussing continuous time systems within our framework. In that case all of the above vector fields are in fact independent of u, and they provide the same information about the system. This is because by the semigroup property of flows it holds that $f_{u+v} = f_u \circ f_v = f_v \circ f_u$, so that $X_u^+ = -X_u^- = Z = -Y_u^+ = Y_u^-$. These equalities help us to understand why the continuous time theory is considerably simpler than the discrete one.

Note that applying these definitions to the inverse system (3) instead of system (1) gives the same vector fields except that the pluses are changed for minuses and vice versa.

Given a vector field Y and a control value u, we can define another vector field from Y by applying a change of coordinates given by the diffeomorphism f_u,

$$(\mathrm{Ad}_u\, Y)(x) = (\mathrm{d}f_u(x))^{-1} Y(f_u(x)).$$

Here $\mathrm{d}f_u$ stands for the differential of f_u with respect to x. Using the diffeomorphisms (4), we may also define

$$(\mathrm{Ad}_{u_k \cdots u_1}\, Y)(x) = (\mathrm{d}f_{u_k \cdots u_1}(x))^{-1} Y(f_{u_k \cdots u_1}(x)),$$

and, applying the even more general family of diffeomorphisms (5),

$$(\mathrm{Ad}_{u_k \cdots u_1}^{\varepsilon_k \cdots \varepsilon_1}\, Y)(x) = (\mathrm{d}f_{u_k \cdots u_1}^{\varepsilon_k \cdots \varepsilon_1}(x))^{-1} Y(f_{u_k \cdots u_1}^{\varepsilon_k \cdots \varepsilon_1}(x)). \tag{6}$$

Clearly, the operators "Ad" so defined are linear operators acting on vector fields Y, and we have that

$$\mathrm{Ad}_{u_k \cdots u_1}^{\varepsilon_k \cdots \varepsilon_1}\, Y = \mathrm{Ad}_{u_1}^{\varepsilon_1} \cdots \mathrm{Ad}_{u_k}^{\varepsilon_k}\, Y. \tag{7}$$

(Note the reversal of indices.) We will use the abbreviated notation $\mathrm{Ad}_0^k\, Y$ for $\mathrm{Ad}_{0 \cdots 0}\, Y$ with $u = 0$ repeated k-times, if $k > 0$, and for $\mathrm{Ad}_{0 \cdots 0}^{-1 \cdots -1}\, Y$, if $k < 0$. Additionally, $\mathrm{Ad}_0^0\, Y = Y$. With this notation we have that

$$(\mathrm{Ad}_0^k\, X_u^+)(x) = \frac{\partial}{\partial v}\bigg|_{v=0} f_0^{-k} \circ f_u^{-1} \circ f_{u+v} \circ f_0^k(x)$$

(see Fig. 5) and, more generally,

$$(\mathrm{Ad}_{u_k \cdots u_1} X_{u_0}^+)(x) = \frac{\partial}{\partial v}\bigg|_{v=0} f_{u_k \cdots u_1}^{-1} \circ f_{u_0}^{-1} \circ f_{u_0+v} \circ f_{u_k \cdots u_1}(x).$$

Since our system is assumed to be invertible, we could apply all definitions to the inverse system (3) instead of (1). Then all the pluses in the superscripts change for minuses and Ad_u changes for Ad_u^{-1}, and vice versa. Therefore, we will have the following fact, which we shall use repeatedly.

REVERSION PRINCIPLE. *Any general property of systems of the type* (1) *that can be expressed in terms of the above defined vector fields is preserved if we change the pluses in the superscripts for the minuses and each* Ad_u *for* Ad_u^{-1}, *and vice versa.*

Remark 3.1. Some of the above defined vector fields can be equivalently defined as follows:

$$X_u^+(x) = (\mathrm{d}f_u(x))^{-1} \frac{\partial}{\partial u} f_u(x),$$

$$(\mathrm{Ad}_{u_k \cdots u_1} X_{u_0}^+)(x) = (\mathrm{d}f_{u_k \cdots u_1}(x))^{-1} X_{u_0}^+(f_{u_k \cdots u_1}(x)).$$

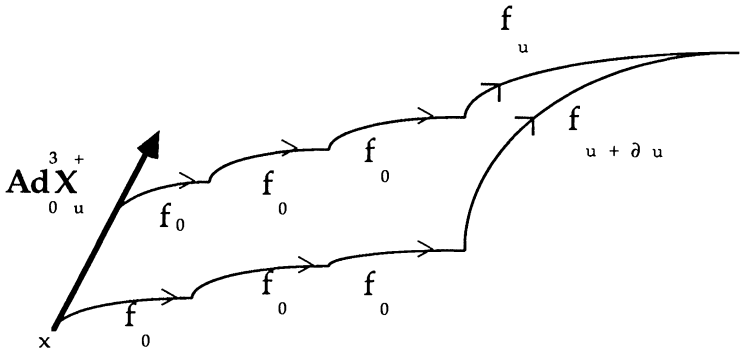


FIG. 5. $\mathrm{Ad}_0^3 X_u^+$.

Since the inverses f_u^{-1} do not appear, the right-hand sides now make sense for *locally* invertible systems. Those of our results that can be stated exclusively in terms of the above vector fields will also hold for locally invertible systems. Furthermore, criteria stated in their terms can be checked without computing the inverse of any diffeomorphism; only matrix inversions are required. For instance, take the system with $\mathbb{X} = \mathbb{R}$, $\mathbb{U} = [-1, 1]$, and equations

$$x^+ = x^3 + 2x + u \sin x.$$

Since for each fixed value of u the right-hand side is strictly increasing, this is an invertible system. We obtain here that

$$X_u^+(x) = \frac{\sin x}{3x^2 + 2 + u \cos x}$$

in the natural coordinates.

The basic interrelations between the vector fields X_u^+, X_u^-, Y_u^+, Y_u^- are given by the following proposition.

PROPOSITION 3.2. *The following equalities hold for each* $u \in \mathbb{U}$.

(*a*) $X_u^+ = -Y_u^+, \quad X_u^- = -Y_u^-$.

(b) $X_u^+ = -\mathrm{Ad}_u X_u^-, \quad Y_u^+ = -\mathrm{Ad}_u Y_u^-$.

Proof. To prove (a), we differentiate with respect to u the equality

$$f_u^{-1} \circ f_u(x) = x$$

and we get

$$Y_u^+(x) + X_u^+(x) = 0.$$

The second equality in (a) follows from the first by the reversion principle.

On the other hand, differentiating with respect to v the equality

$$f_u^{-1} \circ f_{u+v}(x) = f_u^{-1} \circ f_{u+v} \circ f_u^{-1} \circ f_u(x)$$

we get $X_u^+ = \mathrm{Ad}_u Y_u^-$, which together with (a) gives (b). The proof of the last equality now follows by the reversion principle. □

Later in the paper it will be very useful to have a formula for the derivative with respect to u of a vector field Y transformed by the diffeomorphism f_u. It was noted in [25], [11], [17], [18] that this derivative can be easily expressed via the above introduced vector fields and the Lie bracket; in fact, the next two propositions appear as the first steps in the proof of Theorem 3 on page 26 of [17] and of Lemma 3 in [18].

Here and further we shall use the standard notation $[Y, Z]$ for the Lie bracket of the vector fields Y and Z which, in $\mathbb{R}^n$, is given by $[Y, Z] = \partial Z/\partial x\ Y - \partial Y/\partial x Z$. We also denote $\operatorname{ad} Z(Y) = [Z, Y]$ and the kth iteration of the operator $\operatorname{ad} Z$, $\operatorname{ad}^k Z(Y) = \operatorname{ad} Z \cdots \operatorname{ad} Z(Y)$. The flow of the vector field Y is denoted by $\exp(tY)$.

PROPOSITION 3.3. *The following equalities hold for any vector field Z and any $u \in \mathbb{U}$:*

$$\frac{\partial}{\partial u} \operatorname{Ad}_u Z = \operatorname{ad} X_u^+(\operatorname{Ad}_u Z)$$

and

$$\frac{\partial}{\partial u} \operatorname{Ad}_u^{-1} Z = \operatorname{ad} X_u^-(\operatorname{Ad}_u^{-1} Z).$$

Proof. It is enough to prove each of the equalities locally, so we shall assume that we are in $\mathbb{R}^n$. We have that

$$\begin{aligned}
\frac{\partial}{\partial u} \operatorname{Ad}_u Z &= \frac{\partial}{\partial t}\bigg|_{t=0} \frac{\partial}{\partial u} f_u^{-1} \circ \exp(tZ) \circ f_u(x) \\
&= \frac{\partial}{\partial t}\bigg|_{t=0} \left(\frac{\partial}{\partial u} f_u^{-1}\right) \circ f_u \circ f_u^{-1} \circ \exp(tZ) \circ f_u(x) \\
&\quad + \frac{\partial}{\partial t}\bigg|_{t=0} \mathrm{d}(f_u^{-1} \circ \exp(tZ) \circ f_u)(x)(\mathrm{d}f_u(x))^{-1} \frac{\partial}{\partial u} f_u(x) \\
&= (\partial Y_u^+/\partial x)(x)\ \operatorname{Ad}_u Z(x) + (\partial \operatorname{Ad}_u Z/\partial x)(x) X_u^+(x) \\
&= [X_u^+, \operatorname{Ad}_u Z](x),
\end{aligned}$$

where we use the equality $X^+ = -Y^+$.

The second equality follows from the first by the reversion principle, replacing f_u by f_u^{-1}. □

In the next proposition and in the rest of the paper we shall use the following notational convention. Given a family of vector fields $\{Y_\alpha | \alpha \in A\}$, we denote by $\operatorname{Lie}\{Y_\alpha | \alpha \in A\}$ the Lie algebra generated by this family of vector fields and by $\operatorname{Lie}\{Y_\alpha | \alpha \in A\}(x)$ the subspace of the tangent space at x generated by the vector fields in this Lie algebra.

PROPOSITION 3.4. *For analytic systems and connected $\mathbb{U}$,*

$$\operatorname{Ad}_0^k X_u^+(x) \in \operatorname{Lie}\{\operatorname{Ad}_0^{k+1} X_u^- | u \in \mathbb{U}\}(x)$$

and

$$\operatorname{Ad}_0^{-k} X_u^-(x) \in \operatorname{Lie}\{\operatorname{Ad}_0^{-k-1} X_u^+ | u \in \mathbb{U}\}(x)$$

for each $x \in \mathbb{X}$, each $u \in \mathbb{U}$, and each integer k.

In the proof of this proposition we shall use the following lemma. This lemma is in fact about identities on free Lie algebras; we give a somewhat informal statement to avoid having to introduce considerably more machinery.

LEMMA 3.5. *For any $r \geqq 0$ there are coefficients $a_1, \cdots, a_p \in \mathbb{R}$ and $b_1, \cdots, b_q \in \mathbb{R}$ independent of x and u such that*

$$\mathrm{Ad}_u \left(\frac{\partial^r}{\partial u^r} X_u^- \right) = \sum a_i Y_i \text{ where } Y_i \in \mathrm{Lie} \left\{ X_u^+, \frac{\partial}{\partial u} X_u^+, \cdots, \frac{\partial^r}{\partial u^r} X_u^+ \right\} = M_u^{r,+},$$

$$\mathrm{Ad}_u^{-1} \left(\frac{\partial^r}{\partial u^r} X_u^+ \right) = \sum b_i Z_i \text{ where } Z_i \in \mathrm{Lie} \left\{ X_u^-, \frac{\partial}{\partial u} X_u^- \cdots, \frac{\partial^r}{\partial u^r} X_u^- \right\} = M_u^{r,-}.$$

Moreover, these coefficients, as well as the expressions of each Z_i and Y_i in terms of the generators of the corresponding Lie algebra of vector fields, are independent of the particular system.

Proof. From Proposition 3.2 it follows that the assertions are true for $r = 0$. Assume that the first of them is true for $r = k$. From Proposition 3.3 it follows that

$$\frac{\partial}{\partial u} \mathrm{Ad}_u \frac{\partial^k}{\partial u^k} X_u^- = \mathrm{ad}\, X_u^+ \, \mathrm{Ad}_u \frac{\partial^k}{\partial u^k} X_u^- + \mathrm{Ad}_u \frac{\partial^{k+1}}{\partial u^{k+1}} X_u^-. \tag{8}$$

In general for parametrized vector fields A_u, B_u we have that

$$\frac{\partial}{\partial u} [A_u, B_u] = \left[\frac{\partial}{\partial u} A_u, B_u \right] + \left[A_u, \frac{\partial}{\partial u} B_u \right].$$

Thus it follows from the induction assumption that the left side term in (8) is a linear combination of elements in $M^{k+1,+}$ and so is the first term on the right. Therefore, the second element on the right is a linear combination of elements in $M_u^{k+1,+}$ and the assertion is true for $r = k+1$.

The second part of the proposition follows from the first and the reversion principle. □

Proof of Proposition 3.4. In the proof we shall use the following corollary to the Taylor formula for an analytic, vector valued function g defined on a connected set $\mathbb{U}$ containing the origin: $\mathrm{span}\{g(u) | u \in \mathbb{U}\} = \mathrm{span}\{g^{(i)}(0) | i \geqq 0\}$. We have

$$\begin{aligned} \mathrm{span}\{\mathrm{Ad}_0^k X_u^+, u \in \mathbb{U}\}(x) &= \mathrm{Ad}_0^k \,\mathrm{span} \left\{ \left. \frac{\partial^r}{\partial u^r} \right|_{u=0} X_u^+, r \geqq 0 \right\}(x) \\ &\subset \mathrm{Ad}_0^k \left(\mathrm{Ad}_0 \,\mathrm{Lie} \left\{ \left. \frac{\partial^r}{\partial u^r} X_u^- \right|_{u=0}, r \geqq 0 \right\} \right)(x) \\ &= \mathrm{Lie}\{\mathrm{Ad}_0^{k+1} X_u^- | u \in \mathbb{U}\}(x). \end{aligned}$$

Here the inclusion follows from Lemma 3.5 (apply Ad_u to both sides of the second equation and then evaluate at $u = 0$); the first and the third equality follow from Taylor's formula.

The second assertion of the proposition is a consequence of the first and the reversion principle. □

Note that it is not claimed in Proposition 3.4 that, for instance, X_u^+ is in the Lie algebra generated by the vector fields $\mathrm{Ad}_0 X_u^-$. The statement pertains only to the equality of the associated distributions, that is, of the tangent spaces at each point.

4. Acessibility criteria. To state our criteria we shall need the following families of vector fields:

$$\Gamma^+ = \{\mathrm{Ad}_{u_k \cdots u_1} X_{u_0}^+ | k \geqq 0, u_0, \cdots, u_k \in \mathbb{U}\},$$

$$\Gamma^- = \{\mathrm{Ad}_{u_k \cdots u_1}^{-1 \cdots -1} X_{u_0}^- | k \geqq 0, u_0, \cdots, u_k \in \mathbb{U}\},$$

$$\Gamma = \{\mathrm{Ad}_{u_k \cdots u_1}^{\varepsilon_k \cdots \varepsilon_1} X_{u_0}^{\sigma} | k \geqq 0, u_0, \cdots, u_k \in \mathbb{U}, \varepsilon_1, \cdots, \varepsilon_k = \pm 1, \sigma = \pm\}.$$

As previously, for a family of vector fields Δ, we denote by Lie $\{\Delta\}$ the Lie algebra of vector fields generated by Δ, by $\Delta(x)$ the linear space spanned by the vectors at x given by the vector fields in Δ, and by Lie $\{\Delta\}(x)$ the linear space of tangent vectors at x given by the vector fields in the Lie algebra.

The following theorem gives criteria for accessibility of smooth systems. It will be one of the main results of this paper.

THEOREM 2. *The following properties hold for any smooth system* (1).

(a) *The system is forward accessible if and only if any of the following two equivalent conditions hold*:

$$\dim \Gamma^+(x) = n \quad \forall x \in \mathbb{X}, \quad \textit{or} \quad \dim \operatorname{Lie}\{\Gamma^+\}(x) = n \quad \forall x \in \mathbb{X}.$$

(b) *The system is backward accessible if and only if any of the following two equivalent conditions hold*:

$$\dim \Gamma^-(x) = n \quad \forall x \in \mathbb{X}, \quad \textit{or} \quad \dim \operatorname{Lie}\{\Gamma^-\}(x) = n \quad \forall x \in \mathbb{X}.$$

(c) *The system is transitive if and only if any of the following two equivalent conditions hold*:

$$\dim \Gamma(x) = n \quad \forall x \in \mathbb{X}, \quad \textit{or} \quad \dim \operatorname{Lie}\{\Gamma\}(x) = n \quad \forall x \in \mathbb{X}.$$

To state a stronger version of our result, valid for analytic systems, we need the following Lie algebras of vector fields:

$$L^+ = \operatorname{Lie}\{\operatorname{Ad}_0^k X_u^+ | k \geqq 0,\ u \in \mathbb{U}\},$$

$$L^- = \operatorname{Lie}\{\operatorname{Ad}_0^k X_u^- | k \leqq 0,\ u \in \mathbb{U}\},$$

$$L = \operatorname{Lie}\{\operatorname{Ad}_0^k X_u^\sigma | k \in \mathbb{Z},\ u \in \mathbb{U},\ \sigma \in \{+, -\}\}.$$

The following inclusions are evident:

$$L^+ \subset \operatorname{Lie} \Gamma^+, \quad L^- \subset \operatorname{Lie} \Gamma^-, \quad L \subset \operatorname{Lie} \Gamma.$$

In terms of this data, we now state another one of our main results. As remarked earlier, the transitivity case had been stated before ([11], [20]). Even for that case, however, we believe that this paper contains the first complete proof.

THEOREM 3. *The following properties hold for any analytic system* (1) *with connected* $\mathbb{U}$:

(a) *The system is forward accessible if and only if*

$$\dim L^+(x) = n \quad \textit{for any } x \in \mathbb{X}.$$

(b) *The system is backward accessible if and only if*

$$\dim L^-(x) = n \quad \textit{for any } x \in \mathbb{X}.$$

(c) *The system is transitive if and only if*

$$\dim L(x) = n \quad \textit{for any } x \in \mathbb{X}.$$

Remark 4.1. As a consequence of Proposition 3.4, if we were to take in the definition of the Lie algebra L only $\sigma = +$, or alternatively, only $\sigma = -$, a smaller set of vector fields may result, but the conclusions in the theorem would hold equally well.

There is a pointwise version of the above results. An *equilibrium point* $x_0 \in \mathbb{X}$ is one such that $f(x_0, 0) = 0$.

THEOREM 4. *The following properties hold, if* $\mathbb{U}$ *is connected*:

(a) *A smooth system* (1) *is transitive from* x *if and only if* $\dim \Gamma(x) = n$ (*equivalently*, $\dim \operatorname{Lie}\{\Gamma\}(x) = n$). *An analytic system* (1) *is transitive from* x *if and only if* $\dim L(x) = n$.

(b) *An analytic system* (1) *is forward* (*respectively, backward*) *accessible from an equilibrium point* x_0 *if and only if* $\dim L^+(x_0) = n$ (*respectively,* $\dim L^-(x_0) = n$).

The proofs of all these results are given later after we develop some further theory.

The second part of Theorem 4 will be strengthened as a consequence of the following proposition.

PROPOSITION 4.2. *If the system is analytic,* $\mathbb{U}$ *is connected, and* x_0 *is an equilibrium point, then*

$$L^+(x_0) = L^-(x_0) = L(x_0).$$

Proof. Since $L^+(x_0) \subset L(x_0)$, it is enough to show that $L^+(x_0)$ has the same dimension as $L(x_0)$ to conclude that they are equal. Pick a basis of the latter and assume that the elements in the basis involve vector fields of the form $\mathrm{Ad}_0^k X_u^+$, with the possible k bounded below by the integer k^*. (Recall Remark 4.1 to the effect that we may always assume that $\sigma = +$ in the definition of L.) Applying the operator

$$\mathrm{Ad}_0^{-k^*}$$

to these vector fields, we obtain vector fields in L^+. As x_0 is an equilibrium point, the operator $\mathrm{Ad}_0^{-k^*}$ preserves the tangent space at x_0 and we obtain a set of linearly independent vectors in $L^+(x_0)$, as desired. The argument for L^- follows by the reversion principle. □

The above theorem and proposition immediately imply the following corollary.

COROLLARY 4.3. *Assume that the system is analytic,* $\mathbb{U}$ *is connected, and* x_0 *is an equilibrium point. Then forward accessibility from* x_0, *backward accessibility from* x_0, *and transitivity from* x_0 *are all equivalent properties.*

We will prove the above theorems by splitting them into (somewhat stronger) sufficiency and necessity results.

Define the following families of vector fields:

$$X_u^{+,i} = \frac{\partial^i}{\partial u^i} X_u^+, \qquad X_u^{-,i} = \frac{\partial^i}{\partial u^i} X_u^-.$$

THEOREM 5. *The following statements hold for any smooth system* (1).

(a) *If*

$$\dim \mathrm{Lie}\,\{\Gamma^+\}(x) = n \quad \textit{for all } x \in \mathbb{X}, \tag{9}$$

then the system is forward accessible.

(b) *If* x_0 *is an equilibrium point and if*

$$\dim \mathrm{Lie}\,\{\mathrm{Ad}_0^k X_0^{+,i} | k \geqq 0,\ i \geqq 0\}(x_0) = n, \tag{10}$$

then the system is forward accessible from x_0.

(c) *The same statements hold for backward accessibility if we replace* Γ^+ *for* Γ^- *and* $X_u^{+,i}$ *for* $X_u^{-,i}$.

Proof. (a) Let us fix an $x \in \mathbb{X}$. Let p and $v_1^*, \cdots, v_p^*$ be such that the rank of the Jacobian of the map

$$(v_1, \cdots, v_p) \mapsto f_{v_p \cdots v_1}(x) \tag{11}$$

is maximal (over all $p \geqq 0$ and $v_1, \cdots, v_p \in \mathbb{U}$) at $v_1^*, \cdots, v_p^*$. Because $\mathbb{U} \subset \mathrm{clos\,int}\,\mathbb{U}$, we may assume that these are in the interior of $\mathbb{U}$. Let W be a neighborhood of $(v_1^*, \cdots, v_p^*)$ on which this rank is maximal and such that the image S of W under the above map is a submanifold. Since $S \subseteq A^+(x)$, it is enough to show that the dimension of S is equal to n, from which it will follow that S is an open subset of $\mathbb{X}$.

We now prove that each vector field of the type $\mathrm{Ad}_{u_k \cdots u_1} X^+_{u_0}$ is tangent to S. It will follow then that all the Lie brackets of these vector fields are tangent to the submanifold S. This, together with assumption (9), will imply that S is of dimension n.

Assume that the vector

$$\mu := (\mathrm{Ad}_{u_k \cdots u_1} X^+_{u_{k+1}})(y)$$

is not tangent to S at $y = f_{v_p \cdots v_1}(x)$, for some $u_1, \cdots, u_{k+1}$ (for convenience we denote u_0 by u_{k+1} now) and some $(v_1, \cdots, v_p) \in W$. Again, we may assume that these are all in the interior of $\mathbb{U}$. Thus

$$\mu = \frac{\partial}{\partial v}\bigg|_{v=0} (f_{u_k \cdots u_1})^{-1} \circ f^{-1}_{u_{k+1}} \circ f_{u_{k+1}+v} \circ f_{u_k \cdots u_1}(y)$$

is not tangent to S and therefore also

$$\frac{\partial}{\partial v}\bigg|_{v=0} f_{u_{k+1}+v} \circ f_{u_k \cdots u_1} \circ f_{v_p \cdots v_1} = (\mathrm{d} f_{u_{k+1} \cdots u_1})(y)\mu$$

is not tangent to the submanifold $f_{u_{k+1} \cdots u_1}(S)$. But this means that the rank of the Jacobian map of the mapping

$$(v_1, \cdots, v_p, u_1, \cdots, u_{k+1}) \to f_{u_{k+1} \cdots u_1 v_p \cdots v_1}(x)$$

is at least dim $S+1$ for this sequence $v_1, \cdots, v_p, u_1, \cdots, u_{k+1}$, contradicting maximality of the rank. It follows that the vector field $\mathrm{Ad}_{u_k \cdots u_1} X^+_{u_{k+1}}$ must indeed be tangent to S.

(b) The idea of this part of the proof is the same as in part (a) except that now the rank assumption is made at one point only. Thus, we have to construct the manifold S in a neighborhood of x_0 so that n linearly independent vector fields in the Lie algebra (10) are linearly independent in this neighborhood and tangent to this manifold.

Let V be a coordinate neighborhood of x_0 such that there are n vector fields in the Lie algebra (10) which are linearly independent on V. Suppose that these vector fields involve only $k \leqq k^*$. Let $V_\varepsilon \subset V$ denote the open ball of radius ε centered at x_0. Fix δ so that $V_\delta \subset V$ and denote by r_ε the supremum of the possible ranks of those maps (11) with $p \geqq 1$ and $x = x_0$ for which all the points of the trajectory

$$x_i = f_{v_i \cdots v_1}(x_0), \qquad i = 1, \cdots, p,$$

lie in V_ε. Note that r_ε is nondecreasing with ε. Let $r = \inf\{r_\varepsilon | 0 < \varepsilon < \delta\}$ and let $\varepsilon^* := \sup\{\varepsilon | r = r_\varepsilon\}$. Note that $\varepsilon^* > 0$. Take $0 < \sigma < \varepsilon^*$ such that all trajectories starting from V_σ stay in V_{ε^*} for the next k^*+1 steps, under the constant control $u = 0$. Let the corresponding supremum of ranks defining $r_\sigma = r$ be achieved at p and $(v_1^*, \cdots, v_p^*)$.

We define our manifold S as previously, where W is a neighborhood of $(v_1^*, \cdots, v_p^*)$ such that all trajectories corresponding to controls in W lie in V_σ. By an analogous argument as for (a) we see that the vector fields $\mathrm{Ad}_{u_k \cdots u_1} X^+_{u_0}$ are tangent to S, provided that $k \leqq k^*$ and $u_0, \cdots, u_k$ are close enough to zero so that our trajectory does not leave V_{ε^*}, and so the rank cannot increase over r (cf. the definition of σ). Taking $u_1 = \cdots = u_k = 0$ and the derivative $(\partial/\partial u_0)^i$ at u^* we conclude that the vector fields $\mathrm{Ad}^k_0 X^{+,i}_0$ are tangent to S. Therefore, their Lie brackets must be tangent to S, also. Because of our choice of the neighborhoods, there are n linearly independent vector fields among those Lie brackets and so S is an open subset of $\mathbb{X}$.

Statement (c) follows from (a) and (b) and the reversion principle. □

The above proof, part (a), gives a somewhat stronger result, actually, which we state below for further use.

COROLLARY 4.4. *If $y \in \mathbb{X}$ is a point forward reachable from x with maximal rank (in the sense of the ranks of maps* (11)*), then the condition* $\dim \operatorname{Lie}\{\Gamma^+\}(y) = n$ *implies that the system* (1) *is forward accessible from x.*

We are now ready to establish a converse to Theorem 5.

THEOREM 6. (a) *If system* (1) *is of class C^1 and forward accessible from x, then*

$$\dim \Gamma^+(x) = n.$$

(b) *If system* (1) *is analytic, forward accessible from x, and $\mathbb{U}$ is connected, then*

$$\dim L^+(x) = n.$$

(c) *Analogous results hold for backward accessibility with Γ^+, L^+ replaced by Γ^-, L^-.*

Remark 4.5. The case when $\mathbb{U}$ is a nonconnected subset of $\mathbb{R}$ can also be treated. Assume that $\mathbb{U}$ is a disjoint union of connected subsets of $\mathbb{R}$, each of which is in the closure of its interior. Then (b) also holds but we have to choose a subset $\mathbb{U}_0 \subset \mathbb{U}$ which has at least one point in each of these sets. Then

$$L^+ = \operatorname{Lie}\{\operatorname{Ad}_{u_k \cdots u_1} X_u^+ | k \geqq 0,\ u \in \mathbb{U},\ u_1, \cdots, u_k \in \mathbb{U}_0\}$$

must be used in this case as the definition of L^+.

Proof of Theorem 6. (a) If the system is accessible, then it follows from Proposition 2.3 that, for some $k \geqq 1$ the rank of the map $\psi_{k,x}$ is equal to n at some point. This means that the following vectors span an n-dimensional space, for some sequence $u_1, \cdots, u_k$:

$$\frac{\partial}{\partial u_i} f_{u_k \cdots u_1}(x), \qquad i = 1, \cdots, k.$$

Hence, also the vectors

$$(\mathrm{d} f_{u_k \cdots u_1}(x))^{-1} \frac{\partial}{\partial u_i} f_{u_k \cdots u_1}(x)$$

which can be equivalently written as

$$\frac{\partial}{\partial v}\bigg|_{v=0} f^{-1}_{u_{i-1} \cdots u_1} \circ f^{-1}_{u_i} \circ f_{u_i + v} \circ f_{u_{i-1} \cdots u_1}(x) = \operatorname{Ad}_{u_{i-1} \cdots u_1} X^+_{u_i},$$

$i = 1, \cdots, k$, span an n-dimensional space and statement (a) follows.

(b) The proof will be based on a reduction to continuous time systems, as done in [29] for the transitivity problem. A different proof, not involving such a reduction, is provided in a later section. If our system is accessible from x, then it follows from Proposition 2.3 that there exists a k such that the rank of the map

$$(u_1, \cdots, u_k) \mapsto f_{u_k \cdots u_1}(x)$$

is equal to n at some point $(u_1^*, \cdots, u_k^*)$, and so its image contains an open set V. Then $W = f_0^{-k}(V)$ is also open and $x \in W$. We will show that W is contained in the orbit through x of the Lie algebra L^+ (cf. [34]), which we denote by $\operatorname{Orb}_{L^+}(x)$. This will imply that the orbit is of dimension n and from a theorem of Nagano ([23], [34]) it will follow that $\dim L^+(x) = n$.

Let $y \in W$. We will show that $y \in \operatorname{Orb}_{L^+}(x)$ by showing the equivalent fact: $x \in \operatorname{Orb}_{L^+}(y)$. We have that

$$x = f^{-1}_{u_1} \circ \cdots \circ f^{-1}_{u_k} \circ f_0^k(y) = g_{1,u_1} \circ \cdots \circ g_{k,u_k}(y),$$

where

$$g_{i,u_i} = f_0^{-i+1} \circ f_{u_i}^{-1} \circ f_0^i.$$

Denote $y_k = y$, and

$$y_{i-1} = g_{i,u_i}(y_i), \qquad i = k, \cdots, 1.$$

We have that $y_0 = x$. It is enough to show that $y_{i-1} \in \mathrm{Orb}_{L^+}(y_i)$, for $i = 1, \cdots, k$.

Denote

$$\gamma(u) = f_0^{-i+1} \circ f_u^{-1} \circ f_0^i(y_i).$$

Then, for $u \in [0, u_i]$, γ is a curve in $\mathbb{X}$ joining y_i with y_{i-1}; its tangent vector at u is

$$\frac{\partial}{\partial u}\gamma(u) = \frac{\partial}{\partial v}\bigg|_{v=0} f_0^{-i+1} \circ f_{u+v}^{-1} \circ f_u \circ f_0^{i-1}(\gamma(u)) = \mathrm{Ad}_0^{i-1} Y_u^+(\gamma(u)).$$

As $\gamma(0) = y_i$ and $\mathbb{U}$ is connected, it follows that $y_{i-1} = \gamma(u_i)$ belongs to the orbit through y_i of the family of vector fields $\mathrm{Ad}_0^{i-1} Y_u^+$, $u \in \mathbb{U}$. Since $Y_u^+ = -X_u^+$, it then follows that y_{i-1} belongs to the orbit through y_i of the family $\mathrm{Ad}_0^{i-1} X_u^+$, $u \in \mathbb{U}$. □

Remark 4.6. If $\mathbb{U}$ is not connected, then the result still holds with the modified definition of the Lie algebra L^+ as given in the remark following Theorem 6. The necessary modifications in the above proof are as follows. We choose elements $v_1, \cdots, v_k \in \mathbb{U}_0$ so that v_i belongs to the same connected component of $\mathbb{U}$ as u_i^*. Then we define

$$W = f_{v_k}^{-1} \circ \cdots \circ f_{v_1}^{-1}(V).$$

Then we have that

$$x = g_{1,u_1} \circ \cdots \circ g_{k,u_k}(y), \quad g_{i,u_i} = f_{v_{i-1}\cdots v_1}^{-1} \circ f_{u_i}^{-1} \circ f_{v_i} \circ f_{v_{i-1}\cdots v_1}.$$

Finally, we take the curve

$$\gamma(u) = f_{v_{i-1}\cdots v_1}^{-1} \circ f_u^{-1} \circ f_{v_i \cdots v_1}(y_i),$$

with u in the interval joining u_i and v_i. Differentiation with respect to u now gives the vector fields in the modified Lie algebra L^+ as defined in the remark following Theorem 6.

To obtain criteria for transitivity using Theorems 5 and 6, we may apply the following trick which reduces the transitivity problem to the forward accessibility problem.

Define $\mathbb{U}^\pm$ as the disjoint union of two copies of $\mathbb{U}$ denoted by $\mathbb{U}^+$ and $\mathbb{U}^-$. Consider a system

$$x^+ = f^\pm(x, u), \quad x(t) \in \mathbb{X}, \quad u(t) \in \mathbb{U}^\pm = \mathbb{U}^+ \cup \mathbb{U}^- \tag{12}$$

where $f^\pm(x, u) = f(x, u)$ if $u \in \mathbb{U}^+$ and $f^\pm(x, u) = f^-(x, u) = f_u^{-1}(x)$ if $u \in \mathbb{U}^-$. As the control set $\mathbb{U}^\pm$ has two components, we define its Lie algebra of our new system L^+ using the definition in Remark 4.6 with $\mathbb{U}_0 = \{0^+, 0^-\}$, where $0^+ \in \mathbb{U}^+$ and $0^- \in \mathbb{U}^-$ are two copies of $0 \in \mathbb{U}$. Of course, there is no difficulty in embedding the new control set again in the reals. The following proposition is then clear.

PROPOSITION 4.7. (a) *The Lie algebra L^+ of the system* (12) *is equal to the Lie algebra L of the original system* (1).

(b) *The family of vector fields* Γ^+ *for system* (12) *is equal to the family* Γ *defined by system* (1).

(c) *The forward accessible set of system* (12) *is equal to the orbit of system* (1).

We may now complete the proofs of all the theorems in this section.

Proof of Theorem 2. Statement (a) follows immediately from Theorems 5 and 6, part (a). Statement (b) follows analogously from part (c) of these theorems. Finally, statement (c) is the consequence of statement (a) via the above reduction of the transitivity problem to the forward accessibility problem and Proposition 4.7. □

Proof of Theorem 3. Statement (a) follows from Theorem 5 (a) and the inclusion $L^+ \subset \text{Lie}\{\Gamma\}$ (sufficiency), and from Theorem 6(b). Statement (b) follows analogously from statements (c) of these theorems. Finally, statement (c) is the consequence of statement (a) via the above reduction trick and Proposition 4.7. □

Proof of Theorem 4. (a) In the smooth case the "if" part follows from Corollary 4.4 by the above reduction procedure and Proposition 4.7 as, for system (12) the point x is attainable from itself with full rank. The analytic case follows from the smooth case by the inclusion $L(x) \subset \text{Lie}\{\Gamma\}(x)$.

The "only if" part follows from Theorem 6 and Proposition 4.7 via the above reduction.

(b) The "only if" part is the consequence of Theorem 6. To prove the "if" part suppose that there are n linearly independent vectors in $L^+(x_0)$. Each of them can be taken in the form

$$\text{ad}\,(\text{Ad}_0^{k_1} X_{u_1}^+) \cdots \text{ad}\,(\text{Ad}_0^{k_{p-1}} X_{u_{p-1}}^+)(\text{Ad}_0^{k_p} X_{u_p}^+)(x_0). \tag{13}$$

If we take the partial derivatives of these vectors with respect to $u_1, \cdots, u_p$ at zero, we obtain vectors which appear in the Lie algebra in (10). From the Taylor formula it follows then that the rank condition in (10) is also satisfied and Theorem 5 implies the result. □

5. Nonaccessible systems. In this section we will briefly discuss nonaccessible and, more generally, nontransitive systems. The following "orbit theorem" is crucial in understanding such systems. The theorem has a long history starting with results of Chow, Nagano [23], Sussmann [34], and Stefan [33] in the continuous time case. In the discrete time case, analogous results to those in continuous time were provided in [9], [32], [11], and [29], the latter containing also a proof of a more abstract result dealing with a general notion of action on manifolds. These papers should be consulted for details of the proof, which we omit.

THEOREM 7. *Any orbit* $A(x)$ *of the smooth system* (1) *is an immersed submanifold of* $\mathbb{X}$ *with at most countably many connected components, whose tangent space is given by*

$$T_y A(y) = \Gamma(y)$$

at each $y \in A(x)$. *In the analytic case we have that*

$$T_y A(y) = L(y)$$

holds also.

As the attainable set from x lies in the orbit from x, there is no chance for forward or backward accessibility from x if there is no transitivity from x (that is, the orbit is not of full dimension). In this case it is reasonable to ask whether the attainable set has a nonempty interior in the orbit. In the case of analytic continuous time systems the answer is always positive, as proved by Sussmann and Jurdjevic [36]. The following theorem generalizes this result to discrete time systems.

THEOREM 8. *If x_0 is an equilibrium point of an analytic system* (1), *then each of the attainable sets $A^+(x_0)$ and $A^-(x_0)$ has a nonempty interior in the orbit $A(x_0)$.*

Proof. If we restrict our system to the orbit then the problem reduces to proving that the system is forward (backward) accessible from x_0, if it is transitive from x_0. But this follows immediately from Theorem 4 and Proposition 4.2. □

Remark 5.1. The above theorem provides an analogue of what is sometimes called the *positive form of Chow's lemma* for continuous time systems. In fact, the proof is related to that of the continuous time case. However, there is an interesting subtlety that appears here. Contrary to the continuous situation, it is *not* true now that the assumption that x_0 is an equilibrium state can be relaxed. In the paper [29, Remark 9.15], an example is given of an analytic system on $\mathbb{X}=\mathbb{R}$, with $\mathbb{U}=\mathbb{R}$, and a state $x\in\mathbb{X}$ such that $A(x)=\mathbb{X}$, but the system is not forward accessible from this x. In fact, the system in question arises from the sampling of a continuous time system.

We now give the basic outline of how such an example arises. A real-analytic function of one variable

$$g(x)$$

is first constructed, with the property that

$$|g'(x)|\leqq 1 \text{ for all } x\in\mathbb{R}$$

and whose zeros are exactly at the nonnegative integers $0, 1, 2, \cdots$. Now the system is given by equations

$$x^+ = 1 + x + ug(x)$$

with

$$\mathbb{U}=(-1, 1)$$

as control value set. Observe that this system is indeed invertible, since for each fixed u the right-hand side is a strictly increasing function of x. Furthermore, for each x the set

$$\{x, 1+x, 2+x, \cdots\}$$

is included in $A^+(x)$. When x is a nonnegative integer, this is *precisely* $A^+(x)$, while for any other x one can reach an open set in one step, and hence $A^+(x)$ is of dimension 1. Since each nonnegative integer x can be reached from, say, -1, it follows that $A(x)=A(-1)$ has dimension 1, so by connectedness, *the orbit through each point is all of* $\mathbb{X}=\mathbb{R}$, even though $A^+(0)$, $A^+(1), \cdots$ are discrete.

These remarks probably mean that the notion of transitivity is in the discrete time case too weak to be of interest.

The following families of vector fields will help us to better understand the geometry of the attainable sets $A^+(x)$ and $A^-(x)$ and, in particular, to estimate their dimensions. Define

$$\Delta_k^+ = \{\mathrm{Ad}_0^i X_u^+ | 0\leqq i\leqq k-1, u\in\mathbb{U}\}, \qquad L_k^+ = \mathrm{Lie}\,\Delta_k^+,$$

and

$$\Delta_k^- = \{\mathrm{Ad}_0^{-i} X_u^- | 0\leqq i\leqq k-1, u\in\mathbb{U}\}, \qquad L_k^- = \mathrm{Lie}\,\Delta_k^-.$$

For any family of vector fields Δ, let $\mathrm{Orb}_\Delta(x)$ denote the orbit of this family passing through x. This orbit has a natural structure of immersed second countable submanifold ([34], [33]). Further, the orbit of Lie Δ coincides with the orbit of Δ.

PROPOSITION 5.2. *For any smooth system with connected control set* $\mathbb{U}$ *we have that*

$$A_k^+(x) \subset \mathrm{Orb}_{\Delta_k^-}(y), \quad \textit{for any} \quad y \in A_k^+(x),$$

and

$$A_k^-(x) \subset \mathrm{Orb}_{\Delta_k^+}(y), \quad \textit{for any } y \in A_k^-(x).$$

Proof. It is enough to prove the first inclusion as the second will follow from the reversion principle. It is also enough to show this inclusion for any particular y in $A_k^+(x)$, since for any other y this will be implied by the general equality $\mathrm{Orb}_\Delta(y) = \mathrm{Orb}_\Delta(z)$ for any $z \in \mathrm{Orb}_\Delta(y)$. Our argument will be similar to that used in the proof of Theorem 6(b). Take

$$y = f_0^k(x)$$

and

$$z = f_{u_k \cdots u_1}(x) = f_{u_k} \circ \cdots \circ f_{u_1} \circ f_0^{-k}(y).$$

We have to show that $z \in \mathrm{Orb}_{\Delta_k^-}(y)$. The point z can be written in a different way as

$$z = g_{k,u_k} \circ \cdots \circ g_{1,u_1}(y),$$

where

$$g_{i,u} = f_0^{k-i} \circ f_u \circ f_0^{-k+i-1}, \qquad i = 1, \cdots, k.$$

Taking $z_0 = y$, $z_i = g_{i,u_i}(z_{i-1})$, $i = 1, \cdots, k$, it is enough to show that $z_i \in \mathrm{Orb}_{\Delta_k^-}(z_{i-1})$. Consider the curve $\gamma_i(u) = g_{i,u}(z_{i-1})$, which joins z_{i-1} with z_i when $u \in [0, u_i]$. The tangent vectors to this curve are given by

$$\begin{aligned}\frac{\partial}{\partial u}\gamma_i(u) &= \frac{\partial}{\partial v}\bigg|_{v=0} f_0^{k-i} \circ f_{u+v} \circ f_u^{-1} \circ f_0^{i-k}(\gamma_i(u)) \\ &= \mathrm{Ad}_0^{i-k}\, Y_u^-(\gamma_i(u)) = -\mathrm{Ad}_0^{i-k}\, X_u^-(\gamma_i(u)).\end{aligned}$$

As $-k+1 \leqq i-k \leqq 0$, it follows that the above curve lies in the orbit of the family Δ_k^- and the proof is complete. □

From the above proposition we immediately conclude the following necessary conditions for accessibility.

COROLLARY 5.3. *If an analytic system with connected* $\mathbb{U}$ *is forward accessible from* x, *then*

$$\dim L^-(y) = n \quad \textit{for any } y \in A^+(x).$$

Similarly, if it is backward accessible from x, *then*

$$\dim L^+(y) = n \quad \textit{for any } y \in A^-(x).$$

Proof. The first statement follows directly from the first inclusion in Proposition 5.2 and the inclusions

$$\bigcup_{k>0} A_k^+(x) = A^+(x), \quad \mathrm{Orb}_{\Delta_k^-}(x) \subset \mathrm{Orb}_{L^-}(x).$$

The second statement follows analogously. □

We now turn to yet another reason why our Lie algebras of vector fields emerge in studying controllability properties of discrete time systems. We will consider our system in another (time-dependent) system of coordinates. This is basically the same as the "local" dynamics defined in the references [18] and [20] in the context of invariant distributions for nonlinear discrete-time systems.

Consider the usual system $x(t+1)=f(x(t), u(t))$ and introduce the time-dependent change of variables

$$x(t)=f_0^t(z(t)),$$

where f_0^t is the tth power of f_0 (in the sense of composition). In the new coordinates our system becomes time-dependent and takes the form

$$z(t+1)=g(t, z(t), u(t)), \tag{14}$$

where

$$g(t, z, u)=f_0^{-t-1}\circ f_u\circ f_0^t(z).$$

What is simpler about the new system is that it has the "doing nothing" option, as $g(t, \cdot, 0)=\mathrm{id}$. As a consequence, if the control set $\mathbb{U}$ is connected then so are the attainable sets of system (14): $A^+(x)$, $A^-(x)$, and $A(x)$. In that case the next point on the trajectory, $z(t+1)$, can be connected with the previous one, $z(t)$, by the smooth curve $\gamma(u)=g(t, z(t), u)$, where $u\in[0, u(t)]$ if $u(t)>0$ and $u\in[u(t), 0]$ if $u(t)<0$. As

$$\begin{aligned}\partial\gamma/\partial u(u)&=\partial g/\partial u(t, z(t), u)\\&=\frac{\partial}{\partial v}\bigg|_{v=0}(f_0^{-t-1}\circ f_{u+v}\circ f_u^{-1}\circ f_0^{t+1}(\gamma(u)))\\&=\mathrm{Ad}_0^{t+1}\, Y_u^-(\gamma(u)),\end{aligned}$$

we see that the point $z(t)$ lies in the orbit through $z(t+1)$ of the family of vector fields $\mathrm{Ad}_0^{t+1}\, Y_u^-$, $u\in\mathbb{U}$. Since $Y_u^-=-X_u^-$, it follows by induction that for $t\leqq -1$ any point $z(t)$ on a trajectory of system (14) starting from $z(0)$ lies in the orbit through $z(0)$ of the family of vector fields Δ_k^-, where $k=-t$ and so also in the orbit through $z(0)$ of the Lie algebra L_k^-. By the reversion principle, or by the above argument applied for $t>0$, it also follows that any point $z(t)$ of any trajectory of system (14) starting from $z(0)$ lies in the orbit through $z(0)$ of the family of vector fields Δ_k^+, with $k=t$, and so also in the orbit through $z(0)$ of the Lie algebra L_k^+.

Because of our change of coordinates $x(t)=f_0^t(z(t))$ it follows that a point $x(t)$ on any trajectory of the original system (1) starting from x_0, lies in the image under the map f_0^k of the orbit $\mathrm{Orb}_{\Delta_k^+}(x_0)$ if $t=k>0$ (respectively, the image of $\mathrm{Orb}_{\Delta_k^-}(x_0)$, if $t<0$, $k=-t$). Thus, we have the following proposition.

PROPOSITION 5.4. *If the control set $\mathbb{U}$ is connected then, for any $k>0$, we have the inclusions*

$$A_k^+(x)\subset f_0^k(\mathrm{Orb}_{\Delta_k^+}(x))=f_0^k(\mathrm{Orb}_{L_k^+}(x))$$

and

$$A_k^-(x)\subset f_0^{-k}(\mathrm{Orb}_{\Delta_k^-}(x))=f_0^{-k}(\mathrm{Orb}_{L_k^-}(x)).$$

The orbits of discrete time systems can be expressed via the orbits of the Lie algebra L according to the formula

$$A(x)=\bigcup_{k\in\mathbb{Z}} f_0^k(\mathrm{Orb}_L(x)).$$

Proof. The first two inclusions follow from the argument above. It also follows from the above consideration that the vector fields in L are tangent to the orbit $A(x)$ (cf. Theorem 7). Thus, $\mathrm{Orb}_L(x)\subset A(x)$. As the maps f_0^k preserve the orbit $A(x)$ and the family of vector fields L, it follows that the inclusion "$\supset$" holds. On the other

hand, the computation preceding the proposition also shows that any two points which can be joined by a (forward or backward) step of the discrete time system can also be joined by a trajectory of a continuous time system

$$\dot{x} = h(x, u), \quad \text{where } h(x, u) = \mathrm{Ad}_0^k X_u^+(x)$$

and a (forward or backward) jump by f_0. It is well known that each trajectory of a continuous time system lies in a single orbit of this system. It follows then that any trajectory of the above system lies in an orbit of the family of vector fields L, and so the inclusion "$\subset$" follows. □

The relation between the inclusions in Propositions 5.2 and 5.4 can be further clarified by the following relation between the Lie algebras L_k^+ and L_k^-.

PROPOSITION 5.5. *For an analytic system the distributions spanned by the Lie algebras L_k^+ and L_k^- are related by the change of coordinates given by the diffeomorphism f_0^k, i.e.,*

$$(\mathrm{Ad}_0^k L_k^-)(x) = L_k^+(x), \quad \textit{and} \quad (\mathrm{Ad}_0^{-k} L_k^+)(x) = L_k^-(x) \quad \forall x \in \mathbb{X}.$$

Proof. Since the operator Ad_0 is a homomorphism of the Lie algebra of vector fields, it follows that

$$\mathrm{Ad}_0^k L_k^- = \mathrm{Lie}\,\{\mathrm{Ad}_0^i X_u^- | 1 \leqq i \leqq k\}.$$

From Proposition 3.4 it follows that

$$(\mathrm{Ad}_0^i X_u^-)(x) \in \mathrm{Lie}\,\{\mathrm{Ad}_0^{i-1} X_u^+ | u \in \mathbb{U}\}(x) \quad \forall x.$$

Thus, all the vector fields $\mathrm{Ad}_0^i X_u^-$, $i = 1, \cdots, k$ are tangent to the orbit of the Lie algebra L_k^+ and so

$$(\mathrm{Ad}_0^k L_k^-)(x) \subset L_k^+(x) \quad \forall x \in \mathbb{X}. \tag{15}$$

The reversion principle and the above inclusion yield

$$(\mathrm{Ad}_0^{-k} L_k^+)(x) \subset L_k^-(x) \quad \forall x \in \mathbb{X}.$$

Applying the operator Ad_0^k to both sides of the above inclusion gives the converse inclusion to (15) and proves the first equality in the proposition.

The second equality follows from the first and the reversion principle. □

6. Nonscalar controls. All our previous results can be extended, without difficulties, to the case of multidimensional controls. The basic modification needed is that, whenever derivatives with respect to u are used in the scalar control case, partial derivatives with respect to the components of u should be used in the multicontrol case.

We assume that the control set $\mathbb{U}$ is a subset of $\mathbb{R}^m$ and satisfies the assumption $\mathbb{U} \subset \mathrm{clos\,int}\,\mathbb{U}$. Additionally, we assume that any two points in the same connected component of $\mathbb{U}$ can be joined by a smooth curve lying entirely in $\mathrm{int}\,\mathbb{U}$ (except of endpoints, possibly). We denote $u = (u^1, \cdots, u^m)$ and $v = (v^1, \cdots, v^m)$.

The vector fields X_u^+ defined at the beginning of § 3 should now be redefined as follows:

$$X_{u,i}^+(x) = \frac{\partial}{\partial v^i}\bigg|_{v=0} f_u^{-1} \circ f_{u+v}(x),$$

one for each $i = 1, \cdots, m$. Analogously, we define $X_{u,i}^-$, $Y_{u,i}^+$ and $Y_{u,i}^-$.

The Lie algebras Γ^+, Γ^- and Γ are now defined as

$$\Gamma^+ = \{\mathrm{Ad}_{u_k \cdots u_1} X_{u_0,i}^+ | k \geqq 0, 1 \leqq i \leqq m, u_0, \cdots, u_k \in \mathbb{U}\},$$

$$\Gamma^- = \{\mathrm{Ad}_{u_k \cdots u_1}^{-1 \cdots -1} X_{u_0,i}^- | k \geqq 0, 1 \leqq i \leqq m, u_0, \cdots, u_k \in \mathbb{U}\},$$

$$\Gamma = \{\mathrm{Ad}_{u_k \cdots u_1}^{\varepsilon_k \cdots \varepsilon_1} X_{u_0,i}^{\sigma} | k \geqq 0, 1 \leqq i \leqq m, u_0, \cdots, u_k \in \mathbb{U}, \varepsilon_1, \cdots, \varepsilon_k = \pm 1, \sigma = \pm\}.$$

We also redefine the Lie algebras L^+, L^-, and L as follows. We choose a subset $\mathbb{U}_0 \subset \mathbb{U}$ which has at least one point in each connected component of $\mathbb{U}$. In particular, if the set $\mathbb{U}$ is connected and $0 \in \mathbb{U}$ we can take $\mathbb{U}_0 = \{0\}$. We define

$$L^+ = \text{Lie}\,\{\text{Ad}_{u_k \cdots u_1} X^+_{u,i} | k \geqq 0,\ 1 \leqq i \leqq m,\ u \in \mathbb{U},\ u_1, \cdots, u_k \in \mathbb{U}_0\},$$

$$L^- = \text{Lie}\,\{\text{Ad}^{-1 \cdots -1}_{u_k \cdots u_1} X^-_{u,i} | k \geqq 0,\ 1 \leqq i \leqq m,\ u \in \mathbb{U},\ u_1, \cdots, u_k \in \mathbb{U}_0\},$$

$$L = \text{Lie}\,\{\text{Ad}^{\varepsilon_k \cdots \varepsilon_1}_{u_k \cdots u_1} X^\sigma_{u,i} | k \geqq 0,\ 1 \leqq i \leqq m,\ u \in \mathbb{U},\ u_1, \cdots, u_k \in \mathbb{U}_0,\ \varepsilon_1, \cdots, \varepsilon_k = \pm 1,\ \sigma = \pm\}.$$

THEOREM 9. *With the above definitions of the Lie algebras Γ^+, Γ^-, Γ, L^+, L^-, and L, all the theorems stated in the preceding two sections remain true.*

The proof of the multicontrol versions are completely analogous to the scalar case. The main modifications needed are the replacement of derivatives with respect to u by partial derivatives with respect to the components of u, and the replacement of parameterizations of curves by u with parameterizations by components of u. We leave the details to the reader.

7. From discrete time to continuous time systems. In this section we have two goals. The first is the description of one manner in which the study of continuous time systems can be reduced to that of discrete time systems. The second is the development of a technique, based on expansions of the previously defined families of vector fields, which gives added power to the use of these vector fields and their associated Lie algebras. As an illustration of the use of this technique, we provide a short proof of part (b) of Theorem 6 which is independent of Nagano's theorem and of the orbit theorem. In this manner, not only does the discrete time theory become independent of continuous time techniques, but in fact it becomes itself a basis for the accessibility theory for the latter, via the reduction also described here.

To show how continuous-time systems can be viewed as a special case of discrete time systems, we consider a continuous-time system of the form

$$\dot{x} = h(x, v), \tag{16}$$

where $x(t) \in \mathbb{X}$ and $v(t) \in V$ is the control. We assume that the controls are piecewise constant (this assumption does not affect the controllability properties of the system we are studying). For the convenience of having all the maps defined everywhere we assume that our system is complete. We introduce the discrete-time system

$$x^+ = f(x, u), \quad x(t) \in \mathbb{X},\ u(t) \in \mathbb{U} = \mathbb{R}_+ \times V, \quad \mathbb{R}_+ = [0, \infty), \tag{17}$$

where $u = (t, v)$ and $f(x, u) = \exp(th(\cdot, v))(x)$. In this way, going forward by time t with a constant control v for the continuous-time system corresponds to a forward step using the control $u = (t, v)$ for the discrete time system. Analogously, going backward by time $-t$ with the control v corresponds to a backward step with $u = (t, v)$. This implies that the forward (respectively, backward) attainable sets as well as the orbits of both systems (16) and (17) coincide. Thus both systems have identical controllability properties.

It is convenient to endow V with the discrete topology. The set $\mathbb{U} = \mathbb{R}_+ \times V$ can be viewed then as the disjoint union of copies of $\mathbb{R}_+$. We compute the Lie algebras L^+, L^-, and L corresponding to system (17) according to the remark following Theorem 6. We choose the subset $\mathbb{U}_0 = \{(0, v) | v \in V\} \subset \mathbb{U}$. Then $f_0 = \text{id}$ and we can easily compute that

$$X^+_u = h(\cdot, v) = -X^-_u, \quad \text{for } u = (t, v).$$

Strictly speaking, the present set $\mathbb{U}$ is not an allowable control set, since it is not a subset of $\mathbb{R}^m$. However, the arguments in previous sections can be repeated as long

as we use in the definition of X_u^+ and X_u^- only differentiation with respect to t but not differentiation with respect to v. Finally, we obtain

$$L^+ = L^- = L = \text{Lie}\{h(\cdot, v) | v \in V\}.$$

Our aim now is to prove a discrete time version of the well-known Baker-Campbell-Hausdorff expansion formula for a vector field Y transformed by the flow of a vector field Z:

$$\text{Ad}_u Y = \sum_{k=0}^{\infty} \frac{1}{k!} \text{ad}^k Z(Y).$$

This is classical when Ad_u corresponds to $f_u = \exp(uZ)$, for which $X_u^+ = Z = -X_u^-$. Assume now that f is of the general form $f = f(x, u)$; we wish to generalize the above formula.

LEMMA 7.1. *For analytic f and Y we have the following expansions, for u sufficiently close to zero,*

$$\text{Ad}_u Y = \sum_{k=0}^{\infty} \int_0^u \int_0^{v_1} \cdots \int_0^{v_{k-1}} \text{ad } X_{v_1}^+ \cdots \text{ad } X_{v_k}^+ \text{Ad}_0 Y \, dv_k \cdots dv_1,$$

$$\text{Ad}_u^{-1} Y = \sum_{k=0}^{\infty} \int_0^u \int_0^{v_1} \cdots \int_0^{v_{k-1}} \text{ad } X_{v_1}^- \cdots \text{ad } X_{v_k}^- \text{Ad}_0^{-1} Y \, dv_k \cdots dv_1,$$

where the series converge pointwise at each $x \in \mathbb{X}$.

If f and Y are of class C^∞ only, then we have the formula

$$\text{Ad}_u^{\pm 1} Y = \sum_{i=0}^{k} \int_0^u \int_0^{v_1} \cdots \int_0^{v_{i-1}} \text{ad } X_{v_1}^{\pm} \cdots \text{ad } X_{v_i}^{\pm} \text{Ad}_0^{\pm 1} Y \, dv_i \cdots dv_1 + R_k, \tag{18}$$

where

$$R_k = \int_0^u \int_0^{v_1} \cdots \int_0^{v_k} \text{ad } X_{v_1}^{\pm} \cdots \text{ad } X_{v_{k+1}}^{\pm} \text{Ad}_{v_{k+1}}^{\pm 1} Y \, dv_{k+1} \cdots dv_1.$$

(Note the subscript "0" in $\text{Ad}_0^{\pm 1} Y$ in each of the above formulas except for the one for the reminder term R_k.)

In order to prove the above lemma we shall first prove the following estimate. Below we shall denote by $|\psi|$ the absolute value of ψ, if ψ is a scalar, and the "max" norm $|\psi| = \max\{|\psi_1|, \cdots, |\psi_n|\}$, if ψ is a vector $\psi = (\psi_1, \cdots, \psi_n)$.

LEMMA 7.2. *Let x be a point in $\mathbb{R}^n$. If $Y_0, \cdots, Y_k$ are real analytic vector fields on a subset of $\mathbb{R}^n$ containing x that have complex analytic continuations (denoted by the same letters) to the closed polydisc $D = D_{x,r} = \{z \in \mathbb{C}^n | |z_1 - x_1| \leqq r, \cdots, |z_n - x_n| \leqq r\}$, then*

$$|\text{ad } Y_k \cdots \text{ad } Y_2(Y_1)(x)| \leqq \sup_{z \in D} |Y_k(z)| \cdots \sup_{z \in D} |Y_1(z)| (2/r)^{k-1} k^k. \tag{19}$$

Proof. Before we prove the estimate in the lemma, we shall derive the following estimate. Let ϕ be a real analytic function which has a complex analytic extension to the polydisc D. Then the iterated derivative of ϕ along the vector fields $Y_1, \cdots, Y_k$ can be estimated by

$$|Y_k \cdots Y_1 \phi(x)| \leqq \sup_{z \in D} |\phi(z)| \sup_{z \in D} |Y_1(z)| \cdots \sup_{z \in D} |Y_k(z)| (k/r)^r. \tag{20}$$

To prove this estimate we use a method of Sussmann [34] (proof of Lemma 4.2) which reduces the problem to Cauchy inequalities. Consider the complex analytic vector

fields $z_1 Y_1, \cdots, z_k Y_k$ defined on D, where $z_1, \cdots, z_k$ are complex parameters in the unit disc $\{z \| z| \leqq 1\}$. Let $\exp(t z_i Y_i)$ denote the flow of $z_i Y_i$ in $\mathbb{C}^n$. Then

$$\phi \circ \exp(t_1 z_1 Y_1) \circ \cdots \circ \exp(t_k z_k Y_k)(x)$$

is a well-defined analytic function on the unit polydisc $|z_1| \leqq 1, \cdots, |z_k| \leqq 1$, if

$$|t_i| \leqq r(k \sup |Y(z)|_{z \in D})^{-1}, \qquad i = 1, \cdots, k \tag{21}$$

(as the concatenation of the trajectories of $z_1 Y_1, \cdots, z_k Y_k$ starting from x does not leave D if $t_1, \cdots, t_k$ satisfy the above inequalities). From the Cauchy inequality we obtain then that the iterated derivative at the origin of this function with respect to $z_1, \cdots, z_k$ is estimated by the supremum of this function on the unit polydisc. This gives the inequality

$$|(t_k Y_k) \cdots (t_1 Y_1)\phi(x)| \leqq \sup_{z \in D} |\phi(z)|.$$

If we take the maximal values of $t_1, \cdots, t_k$ in the inequalities (21), the above gives (20).

The estimate in (20) gives the inequalities

$$|Y_{i_k} \cdots Y_{i_1} \phi_i(x)| \leqq \sup_{z \in D} |Y_1(z)| \cdots \sup_{z \in D} |Y_k(z)| k^k r^{-k+1}, \tag{22}$$

for $\phi_i = x_i$ and $i_1, \cdots, i_k$ any permutation of $1, \cdots, k$. These inequalities imply the estimate in (19) as the left-hand side of this estimate can be replaced by the components of the vector field given by ad $Y_k \cdots$ ad $Y_1 \phi_i$ and each such component consists of 2^{k-1} terms of the form as in (22) (this follows from the definition of the Lie bracket as a commutator). □

Proof of Lemma 7.1. Integration of the first equation in Proposition 3.3 between 0 and u gives

$$\mathrm{Ad}_u Y = \mathrm{Ad}_0 Y + \int_0^u \mathrm{ad}\, X_v^+ (\mathrm{Ad}_v Y)\, \mathrm{d}v.$$

Replacing $\mathrm{Ad}_v Y$ on the right by this expression yields

$$\mathrm{Ad}_u Y = \mathrm{Ad}_0 Y + \int_0^u \mathrm{ad}\, X_v^+ (\mathrm{Ad}_0 Y)\, \mathrm{d}v + \int_0^u \int_0^{v_1} \mathrm{ad}\, X_{v_1}^+ \mathrm{ad}_{v_2}^+ (\mathrm{Ad}_{v_2} Y)\, \mathrm{d}v_2\, \mathrm{d}v_1.$$

Repeating such a replacing k times gives the "+" case of formula (18). The "−" case follows by the reversion principle.

To prove the first formula of the lemma we shall now use the estimate in Lemma 7.2. Our families of vector fields, X_u^+ and $\mathrm{Ad}_u Y$, are analytic with respect to x and u. Let us fix an $x \in \mathbb{X}$. Then, there exist an $r > 0$ and a u_0 such that both families have complex analytic extensions to the complex polydisc D in $\mathbb{C}^n$, with the (real) center at x and radius r, for all $u \in [0, u_0]$. Denote

$$C = \sup_{z \in D, u \in [0, u_0]} |X_u^+(z)|, \qquad D = \sup_{z \in D, u \in [0, u_0]} |\mathrm{Ad}_u Y(z)|.$$

Lemma 7.2 gives the following estimate for $R_k(x)$ with, if $u \in [0, u_0]$,

$$|R_k(x)| \leqq (2/r)^k (k+1)^{k+1} C^{k+1} D \int_0^u \int_0^{v_1} \cdots \int_0^{v_k} \mathrm{d}v_{k+1} \cdots \mathrm{d}v_1$$

$$= C D u_0 (2 C u_0 / r)^k \frac{(k+1)^{k+1}}{(k+1)!}.$$

From Stirling's formula,

$$\lim_{k\to\infty} (2\pi k)^{1/2} e^k k^k (k!)^{-1} = 1,$$

it follows then that $R_k(x)$ tends to zero as k tends to infinity. This implies that the first series in the lemma converges.

The second formula follows from the first by the reversion principle. □

Both expansions in Lemma 7.1 can be combined to obtain a more general expansion. In order to have a compact expression for this expansion we introduce the following notation. Define the following linear operators acting on vector fields Y or, more generally, on smooth families of vector fields $Y_{(\cdot)}$ depending on $u \in \mathbb{U}$:

$$\operatorname{ad} I_u^{\sigma,i} Y_{(\cdot)} := \int_0^u \int_0^{v_1} \cdots \int_0^{v_{i-1}} \operatorname{ad} X_{v_1}^{\sigma} \cdots \operatorname{ad} X_{v_i}^{\sigma} Y_{v_i} \, dv_i \cdots dv_1,$$

and $\operatorname{ad} I_u^{\sigma,0} Y_{(\cdot)} = Y_u$, where σ is either $+$ or $-$. With this notation, formula (18) in Lemma 7.1 takes the form

$$\operatorname{Ad}_u^{\pm 1} Y = \sum_{i=0}^{k} \operatorname{ad} I_u^{\pm,i} \operatorname{Ad}_0^{\pm 1} Y + \operatorname{ad} I_u^{\pm,k+1} \operatorname{Ad}_{(\cdot)}^{\pm 1} Y.$$

Finally, using analogous techniques as above, one can also establish the forward/backward version of the above.

LEMMA 7.3. *If f and the vector field Y are analytic, then the following expansion holds*:

$$\operatorname{Ad}_{u_k \cdots u_1}^{\varepsilon_k \cdots \varepsilon_1} Y = \sum_{i_1 \geqq 0, \cdots, i_k \geqq 0}^{\infty} \operatorname{ad} I_{u_k}^{\sigma_k, i_k} \operatorname{Ad}_0^{\varepsilon_k} \cdots \operatorname{ad} I_{u_1}^{\sigma_1, i_1} \operatorname{Ad}_0^{\varepsilon_1} Y,$$

where σ_j is the sign of ε_j, $j = 1, \cdots, k$, and the series converges pointwise for small enough u's.

From this we can draw the following conclusions.

COROLLARY 7.4. *If the system is analytic and $\mathbb{U}$ is connected, then*

$$L^+(x) = \Gamma^+(x), \quad L^-(x) = \Gamma^-(x), \quad L(x) = \Gamma(x),$$

for any $x \in \mathbb{X}$.

Again, the result is valid also in the nonconnected case provided that one modifies the definitions of the Lie algebras as explained in Remark 4.5.

Because of Corollary 7.4, part (b) is equivalent to part (a) in Theorem 6. This provides the promised direct proof of part (b) of Theorem 6.

8. Sampling. In this section, we explain briefly how some of our results can be applied to the sampling problem. More details are given in the conference paper [31]. For other related facts about sampling, the reader should consult [19] and [21].

When a continuous-time system is digitally controlled, decisions are often restricted to be taken at fixed times $0, \delta, 2\delta, \cdots$; $\delta > 0$ is the *sampling time.* Under what is often called zeroth-order hold sampled control, the resulting situation can be modeled through the constraint that the inputs applied be constant on intervals of length δ. It is thus of interest to characterize the preservation of basic system properties when the controls are so restricted. For controllability, this problem motivated the results in the classical paper of Kalman, Ho, and Narendra [13]. This studied the case of linear systems and established that controllability when sampling at intervals of length δ is preserved if $\delta(\lambda - \mu)$ is not of the form $2k\pi i$ for any pair of distinct eigenvalues of the A matrix. The dual version of this result, for observability, is basically the classical

Nyquist–Shannon sampling theorem from digital signal processing, and is often summarized by the statement that controllability (or observability) is preserved provided that one samples at more than twice the natural frequencies of the system. We sketch here how a similar result can be obtained for certain nonlinear systems, using the accessibility conditions given above. This is an improvement over the result in [30], where only the case of bilinear systems was treated, and more importantly, where only transitivity conditions were obtained.

Let $\tilde{\Sigma}_d$ denote the class of all continuous time systems Σ of the type

$$\dot{x} = Fx + \sum_{i=1}^{m} u_i g_i(x), \tag{23}$$

where F is an n by n matrix and the coordinates of all the g_i are polynomials of degree at most d. For instance, $\tilde{\Sigma}_0$ is the class of all linear systems (the g_i's are constant vectors), while $\tilde{\Sigma}_1$ is the class of bilinear systems. Here $x(t) \in \mathbb{R}^n$ and $u_i(t) \in \mathbb{R}$ for each t; n is the dimension of the system, m the number of independent controls. We shall study controllability properties of (23) from the initial state $x_0 = 0$. Nonequilibrium initial states can also be studied, but we restrict ourselves to the equilibrium case, always reducible to $x_0 = 0$, for simplicity. We let $f(x) = Fx$ be the linear vector field corresponding to the matrix F.

We shall say that the *natural frequencies* of the system (23) are the imaginary parts of the eigenvalues of F, and let $\Omega(\Sigma, 0)$, or just Ω, be the set of these numbers (counted with multiplicities). Note that since F is real, $-\omega \in \Omega$ whenever $\omega \in \Omega$. For each nonnegative integer j we denote by $\mathscr{B}_j$ the set of all linear combinations

$$\frac{1}{k} \sum_{i=1}^{n} \rho_i \omega_i \tag{24}$$

with k any nonzero integer, $\omega_1, \cdots, \omega_n$ the natural frequencies, and the ρ_i's nonnegative integers satisfying

$$\sum_{i=1}^{n} \rho_i = 2j + 2.$$

Note that if λ is the largest of the ω_i (equivalently, the largest absolute value of these), each element of $\mathscr{B}_j$ is in magnitude bounded by $(2j+2)\lambda$.

Denote the set of states of the continuous time system Σ that can be reached from 0 in time $T > 0$, using arbitrary (measurable locally integrable) controls $u(\cdot)$ by A^T. We shall say that the system (23) is (*forward*) *accessible from* 0 if A^T has nonempty interior for some $T > 0$. Let $\omega > 0$ be any real number. We shall say that Σ is *ω-accessible from* 0, or *accessible under sampling at frequency ω from* 0, if the set of states A_ω^T reachable from 0 in time T using controls sampled at that frequency has a nonempty interior. A control $u(\cdot)$ defined on an interval $[0, T]$ is said to be sampled at frequency ω (in radians/sec) if and only if T is an integer multiple of $\delta := 2\pi/\omega$, say $T = r\delta$, and there are vectors

$$v_1, \cdots, v_r$$

such that $u(t) \equiv v_i$ on the interval $[(i-1)\delta, i\delta)$. Thus accessibility under sampling corresponds to forward accessibility for a discrete time system derived from the corresponding Σ and ω. With this definition it is clear that ω-accessibility for even a single ω implies accessibility. The following theorem from [31] provides a converse to this fact. The corollary is immediate from the theorem and the discussion given above about the largest frequency λ.

THEOREM 10. *Assume that $\Sigma \in \tilde{\Sigma}_d$ is accessible from 0. If $\omega > 0$ is not in $\mathscr{B}_j$ for any $j \leqq d$, then Σ is also ω-accessible.*

COROLLARY 8.1. *Accessibility is preserved under sampling for systems in $\tilde{\Sigma}_d$ provided that the sampling frequency be larger than $2d+2$ times the largest natural frequency of the system.*

The reader is referred to [31] for the details of the reduction of the above theorem to the results given earlier in this paper. However, we wish to at least sketch this reduction here. For each fixed δ, the vector fields X^+ can be explicitly described using a Lie expansion formula ([4], see also [25], and especially [19], [21]):

$$X^+_{0,i} = \frac{\partial}{\partial \varepsilon}\bigg|_{\varepsilon=0} e^{-\delta f} e^{\delta(f+\varepsilon g_i)}(x).$$

(We will be interested here only on the case $u = 0$.) Under suitable assumptions, which are satisfied for the class of systems considered here, this can also be written as

$$\theta_\delta(\mathrm{ad}\, f)(g_i),$$

where as earlier $\mathrm{ad}\, f$ is the operator $\mathrm{ad}\, f(h) = [f, h]$ and for each fixed real number δ, θ_δ is the entire function

$$\theta_\delta(z) := \frac{e^{\delta z} - 1}{z}.$$

Finally, one also has a formal expression, for each fixed δ,

$$\mathrm{Ad}_0 = e^{\delta\, \mathrm{ad}\, f}.$$

This expression can be made rigorous when acting on polynomial vector fields such as those that appear in the classes $\tilde{\Sigma}_d$. Thus the Lie algebra L^+, for each fixed δ, contains the Lie algebra $\tilde{L}^+$ generated by the vector fields

$$\{\theta_\delta(\mathrm{ad}\, f)(g_1), \cdots, \theta_\delta(\mathrm{ad}\, f)(g_m), e^{\delta\, \mathrm{ad}\, f}\theta_\delta(\mathrm{ad}\, f)(g_1), \cdots, e^{\delta\, \mathrm{ad}\, f}\theta_\delta(\mathrm{ad}\, f)(g_m), \cdots,$$
$$e^{\delta k\, \mathrm{ad}\, f}\theta_\delta(\mathrm{ad}\, f)(g_1), \cdots e^{\delta k\, \mathrm{ad}\, f}\theta_\delta(\mathrm{ad}\, f)(g_m), \cdots\},$$

which equals the span of the vector fields

$$\{g_1, \cdots, g_m, [f, g_1], \cdots, [f, g_m], \cdots, \mathrm{ad}^k f(g_1), \cdots, \mathrm{ad}^k f(g_m), \cdots\}$$

when δ is as in Theorem 10 (see [31] for details). It follows that $\tilde{L}^+$ coincides with the strong accessibility Lie algebra associated to the original continuous time system, which has full rank at the origin due to the accessibility assumption. Then Theorem 4 gives the desired result.

9. An example. Consider the following invertible polynomial system with $\mathbb{X} = \mathbb{R}^3$.

$$\begin{aligned} x^+ &= x(z^2+1)^2 \\ y^+ &= y(z^2+1)^3 \\ z^+ &= z + u, \end{aligned} \tag{25}$$

where we are using the superscript + to denote time shift, and we denote coordinates as (x, y, z). Calculating, we obtain that $X^+_u = -2z(z^2+1)^{-1}Z - X^-_u$ and $X^-_u = (0, 0, -1)'$, where Z is the vector field

$$\begin{pmatrix} 2x \\ 3y \\ 0 \end{pmatrix},$$

for each $u \in \mathbb{R}$. Since the basic vector fields X_u^+ and X_u^- turn out to be independent of u in this example, we drop the subscripts u from now on. Further,

$$\text{Ad}_0 X^+ = \begin{pmatrix} -8xz(z^2+1)^{-1} \\ -12yz(z^2+1)^{-1} \\ 1 \end{pmatrix} = 2X^+ + X^-, \tag{26}$$

from which it follows that

$$\text{span}\{X^+, \text{Ad}_0 X^+\} = \text{span}\{X^+, X^-\}.$$

The identity $\text{Ad}_0 X^- = -X^+$ (cf. Proposition 3.2(b)) implies that

$$\text{Ad}_0^2 X^+ = 2\,\text{Ad}_0 X^+ - X^- \in \text{span}\{X^+, X^-\},$$

so the linear span of the set of all generators of L^+, $\{\text{Ad}_0^k X^+, k \geqq 0\}$, coincides with the span of X^+ and X^-. Similarly, applying Ad_0^{-1} to both sides of (26),

$$\text{Ad}_0^{-1} X^- = X^+ - 2\,\text{Ad}_0^{-1} X^+ = X^+ + 2X^-,$$

so the span of the $\{\text{Ad}_0^k X^-, k \leqq 0\}$, the generators of L^-, is again the same. Finally,

$$[X^+, X^-] = 2(1-z^2)(z^2+1)^{-2} Z,$$

from which it follows that $\{X^-, X^+, [X^+, X^-]\}$ and $\{X^-, Z\}$ span the same C^∞ submodule of vector fields. The latter set is involutive, and we conclude that, for this example,

$$L^+ = L^- = L.$$

Thus the orbits have dimension 2 through each point except at those points with $x = y = 0$, where Z vanishes, and there the dimension is 1. The tangent spaces are given by the vectors $\partial/\partial z$ and $2x\partial/\partial x + 3y\partial/\partial y$. The forward and backward accessible sets contain open subsets of each orbit, by the equality of these Lie algebras.

Of course, in this very simple example one can analyze the system directly. The initial states (x_0, y_0, z_0) with $x_0 = y_0 = 0$ are such that the only possible directions of movement are those in which z changes, as is clear from the equations (25), consistently with the above conclusion about tangent spaces. The points where exactly one of x_0 or y_0 is nonzero are also easy to analyze. Take now a point with both x_0 and y_0 nonzero. Consider the set C consisting of all points (x, y, z) with

$$y_0^2 x^3 = x_0^3 y^2.$$

This is the cross product of a cusp with a line. The forward accessible set consists of all (x, y, z) in C with sign $y = \text{sign } y_0$ for which $|x| \geqq |x_0|$ and $|y| \geqq |y_0|$. The backward accessible has both these inequalities reversed, and the orbit consists of the branch of C with just sign $y = \text{sign } y_0$. Note how each such set C, an algebraic variety, can be stratified into three submanifolds, which turn out to be its singular set (the orbit of $(0, 0, 0)$), the orbit of (x_0, y_0, z_0), and the orbit of the "conjugate" point $(x_0, -y_0, z_0)$. See Fig. 6 for a picture of a typical cross-section with constant z.

Thus in this example both the forward-accessible set and the orbit from each point are open subsets of an irreducible algebraic variety. More generally, similar behavior may be expected when dealing with invertible polynomial systems and equilibrium initial states. We conjecture that the orbit is an open subset of the *quasi-reachable* set in the sense of [26] and [27]. This is an algebraic variety, and it can be computed explicitly, via Jacobians of the n-step transition map. Note that polynomial invertible systems may exhibit highly nonlinear behavior, such as in the case $x^+ = x^3 + x + u$,

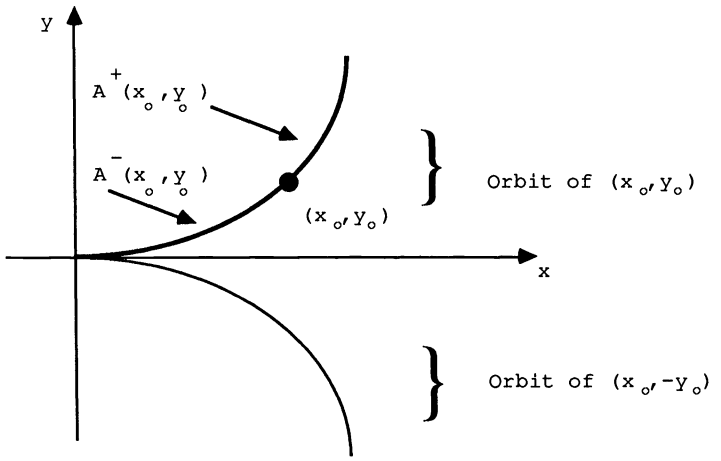


FIG. 6. *Forward and backward accessible sets in example* $(x_0, y_0 > 0)$.

where the inverse of the transition mapping is not even rational. We plan to study such systems in greater depth in the future.

10. An alternative formalism. We now briefly describe the formalism due to Monaco and Normand-Cyrot; the thesis [25] and the papers [17]–[22], as well as the references given there, should be consulted for details.

Their approach is based on the introduction of certain operators and the formal relations that these satisfy. As a first step, one writes the system equations as

$$x^+ = x + f(x, u)$$

so that the new "f" is our $f(x, u) - x$. Thus now f indicates what the increment is, rather than the new state, making things more analogous to differential equations. (This is similar to the introduction of the forward difference operator in numerical analysis.)

For simplicity we shall assume again that inputs are scalar, and also that $\mathbb{X} = \mathbb{R}^n$. Thus we may identify functions $F: \mathbb{R}^n \to \mathbb{R}^n$ (in particular, the functions $F = f(\cdot, u)$) with vector fields, in the usual coordinate system for $\mathbb{R}^n$,

$$F = \sum_{i=1}^{n} F_i(\cdot) \frac{\partial}{\partial x_i}.$$

We will work purely formally, since the intent is merely to point out the relations with the alternative notations in the papers mentioned above. Formally then, one introduces the operators on smooth functions

$$L_F^{\otimes k} := \sum_{i_1, \cdots i_k = 1}^{n} F_{i_1}(\cdot) \cdots F_{i_k}(\cdot) \frac{\partial^k}{\partial x_{i_1} \cdots \partial x_{i_k}}$$

and the complete series

$$\Delta_F := I + \sum_{k \geqq 1} \frac{1}{k!} L_F^{\otimes k}.$$

Now one can obtain similar series for compositions and inverses of the dynamics map.

Further, the vector fields that we use can be expressed then as

$$X_u^+(x) = \frac{\partial}{\partial v}\Big|_{v=0} \Delta_{f(\cdot,u+v)} \circ \Delta^{-1}_{f(\cdot,u)}(\mathrm{Id})|_x,$$

$$X_u^-(x) = \frac{\partial}{\partial v}\Big|_{v=0} \Delta^{-1}_{f(\cdot,u+v)} \circ \Delta_{f(\cdot,u)}(\mathrm{Id})|_x,$$

$$Y_u^+(x) = \frac{\partial}{\partial v}\Big|_{v=0} \Delta_{f(\cdot,u)} \circ \Delta^{-1}_{f(\cdot,u+v)}(\mathrm{Id})|_x,$$

$$Y_u^-(x) = \frac{\partial}{\partial v}\Big|_{v=0} \Delta^{-1}_{f(\cdot,u)} \circ \Delta_{f(\cdot,u+v)}(\mathrm{Id})|_x,$$

$$\mathrm{Ad}_0^k X_u^\sigma(x) = \frac{\partial}{\partial v}\Big|_{v=0} \Delta^k_{f(\cdot,0)}\Delta^\sigma_{f(\cdot,u+v)} \circ \Delta^{-\sigma}_{f(\cdot,u)}\Delta^{-k}_{f(\cdot,0)}(\mathrm{Id})|_x,$$

and many properties of these vector fields can be obtained from the corresponding expansions.

The reader is directed to the above references for details on how these expansions can be very useful in studying, among others, problems of disturbance decoupling, sampling, Volterra expansions, linearization, and realization.

REFERENCES

[1] A. ARAPOSTATHIS, B. JAKUBCZYK, H.-G. LEE, S. I. MARCUS, AND E. D. SONTAG, *The effect of sampling on linear equivalence and feedback linearization*, submitted.

[2] R. BROCKETT, *Nonlinear systems and differential geometry*, Proc. IEEE, 64 (1976), pp. 61-72.

[3] M. FLIESS AND D. NORMAND-CYROT, *A group-theoretic approach to discrete-time nonlinear controllability*, Proc. IEEE Conf. Decision and Control, Albuquerque, NM, 1981.

[4] R. GOODMAN, *Lifting vector fields to nilpotent Lie groups*, J. Math. Pures et Appl., 57 (1978), pp. 77-86.

[5] J. W. GRIZZLE, *Controlled invariance for discrete-time nonlinear systems with an application to the decoupling problem*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 868-874.

[6] J. W. GRIZZLE AND P. V. KOKOTOVIC, *Feedback linearization of sampled-data systems*, IEEE Trans. Automat. Control, to appear.

[7] R. HERMANN AND A. J. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, 22 (1977), pp. 728-740.

[8] A. ISIDORI, *Nonlinear Control Systems: An Introduction*, Springer-Verlag, Berlin, 1985.

[9] B. JAKUBCZYK, *Invertible realizations of nonlinear discrete time systems*, Proc. Princeton Conf. Inf. Sci. and Sys., 1980, pp. 235-239.

[10] ———, *Feedback linearization of discrete-time systems*, Systems Control Lett., 9 (1987), pp. 411-416.

[11] B. JAKUBCZYK AND D. NORMAND-CYROT, *Orbites de pseudo-groupes de diffeomorphismes et commandabilité des systemes non linéaires en temps discret*, C.R. Acad. Sci. Paris, 298-I (1984), pp. 257-260.

[12] B. JAKUBCZYK AND E. D. SONTAG, *The effect of sampling on feedback linearization*, Proc. IEEE Conf. Decision and Control, Los Angeles, Dec. 1987, pp. 1374-1379.

[13] R. E. KALMAN, Y. C. HO, AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contr. Diff. Eqs., 1 (1963), pp. 189-213.

[14] H. G. LEE, A. ARAPOSTATHIS, AND S. I. MARCUS, *On the linearization of discrete time systems*, Internat. J. Control, 45 (1987), pp. 1103-1124.

[15] S. P. MEYN, *Ergodic theorems for discrete time stochastic systems using a generalized stochastic Lyapunov function*, SIAM J. Control Optim., 27 (1989), pp. 1409-1439.

[16] A. MOKKADEM, *Orbites de semi-groupes de morphismes réguliers. Systèmes non linéaires en temps discret.*, Forum Math., to appear.

[17] S. MONACO AND D. NORMAND-CYROT, *Développements founctionels pour les systèmes non linéaires en temps discret*, report from University of Rome, 1984, submitted for publication.

[18] ———, *Invariant distributions for nonlinear discrete-time systems*, Systems Control Lett., 5 (1984), pp. 191-196.

[19] S. MONACO AND D. NORMAND-CYROT, *On the sampling of a linear control system*, Proc. IEEE Conf. Decision and Control, Ft. Lauderdale, Florida, 1985, p. 1095.

[20] ———, *Nonlinear systems in discrete-time*, in Algebraic and Geometric Methods in Control Theory, M. Fliess and M. Hazewinkel, eds., Reidel, Dordrecht, 1986, pp. 411–430.

[21] ———, *A Lie exponential formula for the nonlinear discrete time functional expansion*, in Theory and Applications of Nonlinear Control Systems, C. Byrnes and A. Lindquist, eds., North Holland, Amsterdam, 1986.

[22] ———, *Invariant distributions under sampling*, in Theory and Applications of Nonlinear Control Systems, C. Byrnes and A. Lindquist, eds., North Holland, Amsterdam, 1986.

[23] T. NAGANO, *Linear differential systems with singularities and an application to transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.

[24] H. NIJMEIJER, *Observability of autonomous discrete-time nonlinear systems: a geometric approach*, Internat. J. Control, 36 (1982), pp. 867–874.

[25] D. NORMAND-CYROT, *Théorie et Pratique des Systèmes Non Linéaires en Temps Discret*, Thèse de Docteur d'Etat, Université Paris-Sud, 1983.

[26] E. D. SONTAG AND Y. ROUCHALEAU, *On discrete-time polynomial systems*, J. Nonlinear Analysis, 1 (1976), pp. 55–64.

[27] E. D. SONTAG, *Polynomial Response Maps*, Springer-Verlag, Berlin, New York, 1979.

[28] ———, *A concept of local observability*, Systems Control Lett., 5 (1984), pp. 41–47.

[29] ———, *Orbit theorems and sampling*, in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Fliess and M. Hazewinkel, eds., Reidel, Dordrecht, 1986, pp. 441–486.

[30] ———, *An eigenvalue condition for sampled weak controllability of bilinear systems*, Systems Control Lett., 7 (1986), pp. 313–316.

[31] ———, *A Chow property for sampled bilinear systems*, in Analysis and Control of Nonlinear Systems, C. I. Byrnes and B. Saeks, eds., North Holland, Amsterdam, 1988, pp. 483–492.

[32] E. D. SONTAG AND H. J. SUSSMANN, *Accessibility under sampling*, Proc. IEEE Conf. Decision and Control, Orlando, Florida, Dec. 1982.

[33] P. STEFAN, *Attainable sets are manifolds*, preprint, University of Wales, 1973.

[34] H. J. SUSSMANN, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–188.

[35] ———, *Lie brackets, real analyticity, and geometric control*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhauser, Boston, 1983.

[36] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 34–49, January 1990

002

A CONSTRAINED LEAST SQUARES REGULARIZATION METHOD FOR NONLINEAR ILL-POSED PROBLEMS*

CURTIS R. VOGEL†

Abstract. This paper deals with a method for solving ill-posed, nonlinear Hilbert space operator equations $F(x) = y$. Regularization is obtained by solving a constrained least squares regularization problem

$$\min \|F(x) - y\|^2 \quad \text{subject to } J(x) \leqq \beta^2.$$

β serves as a regularization parameter, and $J(x)$ is a quadratic penalty functional. To robustly and efficiently solve this regularization problem, we apply a trust region method. At each iteration, the quadratic penalty constraint is retained, a Gauss–Newton approximation to the objective functional is taken, and we add a quadratic trust region constraint. The resulting quadratic subproblem is then reformulated as a nonlinear complementarity problem and solved using Newton's method.

This paper applies methods to find approximate solutions to a severely ill-posed nonlinear first kind integral equation arising in geophysics. The method of Generalized Cross Validation (GCV) is used to pick the regularization parameter when random error is present in the discrete data.

Key words. inverse problems, ill-posed problems, regularization, constrained optimization

AMS(MOS) subject classifications. 45G, 49, 65

1. Introduction. In this paper we introduce a constrained least squares regularization method for solving nonlinear ill-posed problems in a Hilbert space setting. Unless otherwise indicated, "$\|\cdot\|$" refers to the appropriate Hilbert space norm. Consider the operator equation

$$F(x) = y, \tag{1.1}$$

where the operator F is nonlinear and maps a separable Hilbert space X into a separable Hilbert space Y. Problem (1.1) is *well-posed* provided: (i) for any $y \in Y$, there exists a solution $x \in X$ for which $F(x) = y$; (ii) the solution x is unique; (iii) the solution x depends continuously on the data y. Otherwise, the problem is *ill-posed.* Examples of ill-posed nonlinear problems include inverse (i.e., parameter estimation) problems for differential equations [1], [6], inverse scattering [7], and nonlinear Fredholm first kind integral equations [10], in which case F is an integral operator of the form

$$F(x)(t) = \int_a^b k(t, \tau, x(\tau))\, d\tau, \qquad a \leqq t \leqq b, \quad x \in X, \tag{1.2}$$

and k is nonlinear in x.

To obtain reasonable approximate solutions, we apply regularization. Regularization methods replace the ill-posed problem with a stabilized problem whose solution depends on a parameter, referred to as the regularization parameter. These methods should have the following features:

1. The regularized problem is well-posed in the sense that a solution exists. (For nonlinear problems, uniqueness and continuous dependence usually cannot be guaranteed.)

* Received by the editors October 20, 1986; accepted for publication March 24, 1989.

† Department of Mathematical Sciences, Montana State University, Bozeman, Montana 59717. This research was partially supported by the National Science Foundation grant DMS-86-02000.

2. One has a "reasonable" means of choosing the regularization parameter, especially when error is present in the data.

3. As the error in the data tends to zero, regularized solutions should converge to the solution of the unperturbed problem, provided the regularization parameter is picked correctly.

Perhaps the most widely used regularization method is the method of Tikhonov Regularization [13] (also known as Regularized Output Least Squares [3], and in statistical circles as the Penalized Likelihood Method [10], [7]) in which one solves the unconstrained minimization problem

$$\min_{x\in X}\{\|F(x)-y\|^2+\alpha J(x)\}. \tag{1.3}$$

Here $\alpha>0$ is the regularization parameter, and $J(x)$ is a penalty functional whose purpose is to stabilize the minimization and provide a priori information about the solution.

In this paper we consider an alternative approach in which regularization is obtained by solving the constrained least squares regularization problem

$$\min_{x\in X}\|F(x)-y\|^2 \quad \text{subject to } J(x)\leqq\beta^2. \tag{1.4}$$

For this method, β is the regularization parameter.

In § 2 we examine the well-posedness of the regularized problem (1.4). We also examine convergence of solutions of (1.4) to a solution of (1.1) as error in the data tends to zero when β is chosen appropriately. In addition, we discuss stability of regularized solutions.

In § 3 we present a nonlinear ill-posed model problem of the form (1.1), (1.2) arising in geophysics. We also discuss the choice of the spaces X and Y for this particular problem, and we show that the assumptions required in § 2 hold for our model problem.

In § 4 we consider the numerical solution of the regularized problem (1.4) when $J(x)$ is a quadratic functional using a trust region method. At each iteration we apply a Gauss-Newton approximation to the object functional $f(x)=\|F(x)-y\|^2$, thus obtaining a quadratic approximation to f. We retain the quadratic regularization constraint and impose an additional quadratic trust region constraint, where the trust region parameter is chosen to decrease the objective function. The resulting quadratic minimization subproblem is diagonalized using the singular value decomposition and then reformulated as a (quadratic) nonlinear complementarity problem. This dual problem in two variables (the Lagrange multipliers for the primal problem) is then solved using Newton's method. The resulting algorithm is robust and quite efficient.

Finally in § 5, we present some numerical results for our algorithm applied to the model problem of § 3. In this section, we also discuss the practical choice of the regularization parameter β when error is present in the discrete data. We apply the method of Generalized Cross Validation to an example where random error is added to the data.

2. Existence and characterization of regularized solutions. The results of the first two theorems below have been obtained by Seidman and Vogel [11] under somewhat more general conditions. We will verify that the assumptions used below are actually satisfied for our model problem in § 3. We first consider existence of regularized solutions.

To simplify notation, we define the objective functional in (1.4),

$$f(x):=\|F(x)-y\|^2,$$

and the constraint set

$$S_\beta := \{x \in X: J(x) \leqq \beta^2\}.$$

THEOREM 2.1. *Let $F: X \to Y$ be weakly continuous, and let the penalty functional $J: X \to R^+ \cup \{0\}$ be weakly lower semicontinuous. Suppose that for each $\gamma \geqq 0$, $\{x \in X: \|x\| \leqq \gamma\} \cap S_\beta$ is weakly compact. Also, suppose $f(x)$ and the penalty functional $J(x)$ are jointly coercive, i.e.,*

$$\lim_{\|x\| \to \infty} [f(x) + J(x)] = \infty.$$

Then problem (1.4) *has a solution.*

Proof. Let $\{x_k\}$ be a minimizing sequence for (1.4). Then by joint coercivity, there exists $\gamma \geqq 0$ such that $\|x_k\| \leqq \gamma$. By the weak compactness assumption, we can extract a subsequence $\{x_{k(j)}\}$ that converges weakly to some x_*. By weak lower semicontinuity of J, $J(x_*) \leqq \liminf J(x_{k(j)}) \leqq \beta^2$. By weak continuity of F and the lower semicontinuity of the Y-norm,

$$\begin{aligned} \|F(x_*) - y\|^2 &\leqq \liminf \|F(x_{k(j)}) - y\|^2 \\ &= \inf \{\|F(x) - y\|^2: x \in S_\beta\}. \end{aligned}$$

□

To obtain a convergence result for perturbed data, we need to assume local uniqueness of the solution to the unperturbed problem (1.1). We also make assumptions concerning solutions to (1.4) with perturbations to the data y and the operator F:

(A1) Let $\bar{y} \in Y$ and let $\bar{x}$ be the unique solution to $F(x) = \bar{y}$ in the region $S_{\bar{\beta}}$, where $\bar{\beta}^2 := J(\bar{x})$.

(A2) Let $y_k \to \bar{y}$ (strong convergence in Y) and $F_k \to F$ in the sense that if x_k converges weakly to x_*, then $F_k(x_k) \to F(x_*)$.

(A3) Let $\beta_k > 0$ and x_k be chosen so that

$$\|F_k(x_k) - y_k\|^2 = \inf \{\|F_k(x) - y_k\|^2: J(x) \leqq \beta_k^2\}, \qquad J(x_k) \leqq \beta_k^2,$$

and suppose $\lim \beta_k = \bar{\beta}$. (Note that such an x_k exists by Theorem 2.1.)

(A4) Suppose J and $f_k(x) := \|F_k(x) - y_k\|^2$ satisfy the coercivity condition

$$\lim_{\|x\| \to \infty} \inf_k [f_k(x) + J(x)] = \infty.$$

(A5) If $\beta_j^2 \to J(x)$, there exists a sequence $\{x_j\}$ for which $J(x_j) \leqq \beta_j^2$ and x_j converges weakly to x.

(A6) If x_k converges weakly to x_* and $J(x_k) \to J(x_*)$, then x_k converges strongly to x_*.

THEOREM 2.2. Under assumptions (A1)–(A6), x_k converges strongly to $\bar{x}$.

Proof. By the coercivity assumption (A4), $\{x_k\}$ is bounded, so we can extract a subsequence $\{x_{k(j)}\}$ converging weakly to some x_*. Since $\beta_{k(j)}^2 \to \bar{\beta}^2 = J(\bar{x})$, by assumption (A5) we can choose a sequence $\{\bar{x}_j\}$ such that $J(\bar{x}_j) \leqq \beta_{k(j)}^2$ and $\bar{x}_j$ converges weakly to $\bar{x}$. Then by the lower semicontinuity of the Y-norm,

$$\begin{aligned} \|F(x_*) - \bar{y}\| &\leqq \liminf \|F_k(x_{k(j)}) - y_{k(j)}\| \\ &= \liminf \{\|F_{k(j)}(x) - y_{k(j)}\|: J(x) \leqq \beta_{k(j)}^2\} \quad \text{by (A3)} \\ &\leqq \lim \|F_{k(j)}(\bar{x}_j) - y_{k(j)}\| \quad \text{by (A5)} \\ &\leqq \lim [\|F_{k(j)}(\bar{x}_j) - F(\bar{x})\| + \|y_{k(j)} - \bar{y}\|]. \end{aligned}$$

By assumption (A2), the last right-hand side goes to 0, so x_* solves $F(x) = \bar{y}$.

By the weak lower semicontinuity of J, $J(x_*) \leqq \liminf J(x_{k(j)}) \leqq \lim \beta_{k(j)}^2 = \bar{\beta}^2$. By the uniqueness assumption (A1), $x_* = \bar{x}$. Moreover, since the above argument can be repeated for any subsequence, x_k itself converges weakly to $\bar{x}$. We next show that $J(x_k) \to \bar{\beta}^2$. The theorem will then follow from assumption (A6).

By assumption (A3), $\limsup J(x_k) \leqq \bar{\beta}^2$. Now suppose $\liminf J(x_k) < \bar{\beta}^2$. Then there exists $\alpha < \beta$ and a subsequence $x_{k(j)}$ with $J(x_{k(j)}) \leqq \alpha^2$. As was done above, we can extract a further subsequence converging to some $\hat{x}$ for which $F(\hat{x}) = \bar{y}$ and $J(\hat{x}) \leqq \alpha^2 < \bar{\beta}^2$. But this contradicts the uniqueness assumption (A1). Hence, $\liminf J(x_k) = \limsup J(x_k) = \bar{\beta}^2 = J(\bar{x})$. □

Remark 2.3. Assumptions (A5) and (A6) hold for many commonly used penalty functionals. For instance, if $J(x) = \|x\|^2$, (A6) is the Afimov-Stekin condition. Assumption (A5) holds if the constraint sets S_{β_j} are closed and convex, in which case we may take x_j for which $\|x_j - x\| = \inf\{\|u - x\|: u \in S_{\beta_j}\}$.

Remark 2.4. Assuming existence but *not* local uniqueness of solutions to $F(x) = \bar{y}$, the above proof shows only the existence of a subsequence which converges weakly to a solution of $F(x) = \bar{y}$.

We next look at a characterization of regularized solutions. We assume that F is twice continuously differentiable with derivatives denoted by $F'(x)$ and $F''(x)$, respectively. Let the superscript "T" denote Hilbert space adjoint. We will also assume a quadratic form for the penalty functional,

$$J(x) = \langle Bx, x \rangle, \tag{2.1}$$

where B is a bounded, self-adjoint positive definite linear operator on X. Recall that $f(x) := \|F(x) - y\|^2$. In addition, we define the constraint functional

$$c(x) := J(x) - \beta^2.$$

THEOREM 2.5. *If x is a solution to* (1.4) *and $\beta > 0$, then there exists $\lambda \in \mathbf{R}$ such that*

$$f'(x) + \lambda c'(x) = 0, \tag{2.2}$$

$$c(x) \leqq 0, \lambda \geqq 0, \tag{2.3}$$

$$\lambda c(x) = 0. \tag{2.4}$$

Proof. See Luenberger [8, p. 249]. The left-hand side of (2.2) is half the gradient of the Lagrangian

$$L_\lambda(x) := f(x) + \lambda c(x).$$

To see that a solution x to (1.4) is a regular point for $c(x)$, take $h = -\frac{1}{2}x$. At such a solution, $c(x) \leqq 0$, and

$$c(x) + \langle c'(x), h \rangle = J(x) - \beta^2 - J(x) = -\beta^2 < 0. \qquad \square$$

In general, solutions to (1.4) need not be locally unique. The following theorem provides conditions for local uniqueness and continuous dependence of local solutions with respect to perturbations in the data y.

THEOREM 2.6. *Let $f_0(x) := \|F(x) - y_0\|^2$ and suppose x_0 is a solution to*

$$\min_{x \in X} f_0(x) \quad \textit{subject to } c(x) \leqq 0.$$

Suppose also that for some $\alpha > 0$, the Hessian of the Lagrangian,

$$L''_\lambda(x_0) := f_0''(x_0) + \lambda c''(x_0) = 2[F''(x_0)^T(F(x_0) - y_0) + F'(x_0)^T F'(x_0) + \lambda B],$$

satisfies

$$\langle L''_\lambda(x_0)s, s \rangle \geqq 2\alpha \|s\|^2 \textit{ whenever } \lambda \langle c'(x_0), s \rangle = 0. \tag{2.5}$$

Then there exists $\tilde{\alpha} > 0$, $\delta > 0$, $r > 0$ *such that* (1.4) *has a solution* x *which is locally unique in some neighborhood* $\|x - x_0\| < \delta$ *whenever* $\|y - y_0\| < r$, *and*

$$\|x - x_0\| \leqq \tilde{\alpha}\sqrt{\|y - y_0\|}. \tag{2.6}$$

Proof. Since F is twice continuously differentiable, so is f, and

$$f_0(x_0 + s) - f_0(x_0) = \langle f'(x_0), s\rangle + \tfrac{1}{2}\langle f''(x_0)s, s\rangle + o(\|s\|^2). \tag{2.7}$$

If $c(x_0) = 0$ and $c(x_0 + s) \leqq 0$, then since $c(x)$ is quadratic,

$$0 \geqq c(x_0 + s) - c(x_0) = \langle c'(x_0), s\rangle + \tfrac{1}{2}\langle c''(x_0)s, s\rangle.$$

Thus

$$f_0(x_0 + s) - f_0(x_0) \geqq \langle L'_\lambda(x_0), s\rangle + \tfrac{1}{2}\langle L''_\lambda(x_0)s, s\rangle + o(\|s\|^2).$$

By Theorem 2.5, $L'_\lambda(x_0) = f'_0(x_0) + \lambda c'(x_0) = 0$, and by (2.5),

$$f_0(x_0 + s) - f_0(x_0) \geqq \alpha\|s\|^2 + o(\|s\|^2).$$

Thus there exists $\delta > 0$ and $\bar{\alpha}$, $0 < \bar{\alpha} \leqq \alpha$, such that

$$f_0(x_0 + s) - f_0(x_0) \geqq \bar{\alpha}\|s\|^2 \tag{2.8}$$

whenever $\|s\| < \delta$. If $c(x_0) < 0$, then $f'(x_0) = 0$, $\lambda = 0$, and (2.8) follows directly from (2.7) and (2.5). Equation (2.6) now follows from Theorems 4 and 6 in the paper by Alt [2] with $\beta = 2$ and y playing the role of w. □

Remark 2.7. We will observe in § 5 that as β becomes large, λ becomes small, and problem (1.4) becomes unstable. To explain this behavior, suppose $\|F''(x_0)\| \leqq M$, $F'(x_0)$ is compact, and X is infinite-dimensional. We can then choose a sequence $\{s_k\}$ such that $\|s_k\| \leqq 1$, $\langle c'(x_0), s_k\rangle = 0$, and s_k converges weakly to 0. Then

$$\langle L''_\lambda(x_0)s_k, s_k\rangle \leqq 2[M\|F(x_0) - y_0\| + \|F'(x_0)s_k\|^2 + \lambda\|B\|],$$

and since $F'(x_0)$ is compact,

$$\limsup \langle L''_\lambda(x_0)s_k, s_k\rangle \leqq 2[M\|F(x_0) - y_0\| + \lambda\|B\|].$$

Consequently, any constant α for which (2.5) holds must satisfy

$$\alpha \leqq M\|F(x_0) - y_0\| + \lambda\|B\|.$$

Then if $\|F(x_0) - y_0\|$ and λ are very small, α is also very small. In Alt's proof [2], the Lipschitz constant $\tilde{\alpha}$ in (2.6) is inversely proportional to $\bar{\alpha} \leqq \alpha$. In this case, the Lipschitz constant $\tilde{\alpha}$ is very large. This is an indication that problem (1.4) is highly unstable.

3. A model problem. In this section we illustrate the results of § 2 with an example. Consider the nonlinear Fredholm first kind integral equation

$$y(t) = F(x)(t) := \int_a^b \log\left[\frac{(t-\tau)^2 + H^2}{(t-\tau)^2 + (H - x(\tau))^2}\right] d\tau, \tag{3.1}$$

where H is a positive parameter. This equation occurs in inverse gravimetry (see [13, p. 15]). The solution $x(\tau)$, $a \leqq \tau \leqq b$, represents the vertical deviation from constant depth H in the location of the boundary of an object buried beneath the surface of the earth. The geometry is shown in Fig. 1. The data $y(t)$, $a \leqq t \leqq b$, represents gravity measurements at the surface of the earth. In practice, observations of $y(t)$ are available at discrete points, but derivatives of y are not available. Thus we take $Y = L^2(a, b)$. We assume the solution vanishes outside the interval $[a, b]$ and is "smooth" in the

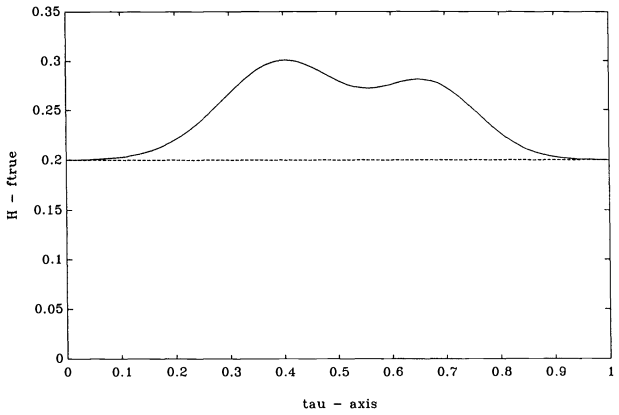


FIG. 1. *Geometry of the gravitational inverse problem.* $z = H - x_{\text{true}}$ *is plotted against* τ.

sense that $\int_a^b x'(\tau)^2\,d\tau$ is bounded. We take $X = H_0^1(a, b) = \{x(\tau), a \leqq \tau \leqq b: x$ is absolutely continuous; $x' \in L^2(a, b)$; $x(a) = x(b) = 0\}$ with inner product

$$\langle x, u\rangle := \int_a^b x'(\tau)u'(\tau)\,d\tau, \qquad x, u \in X, \tag{3.2}$$

and induced norm

$$\|x\| = \sqrt{\int_a^b x'(\tau)^2\,d\tau}, \quad x \in X. \tag{3.3}$$

Remark 3.1. Prilepko [9] has shown that if (3.1) has a solution, the solution is unique provided we make the restriction $H - x(\tau) > 0$. On the other hand, the fact that the right-hand side of (3.1) is analytic as a function of t implies that the range of F is a subset of the analytic functions which is in turn a (dense) proper subset of $Y = L^2(a, b)$. Thus a solution x will not exist for arbitrary $y \in Y$.

The following lemma shows that F is weakly continuous. Since closed balls $\{x \in X: \|x\| \leqq B\}$ are weakly compact, this implies that F is a compact operator. This also shows that the range of F is a proper subset of Y. In addition, the inverse image under F of (noncompact) neighborhoods $N_r(y) = \{z \in Y: \|z - y\| \leqq r\}$, $r > 0$, is unbounded. Hence, we do *not* have continuous dependence of solutions x on the data y.

LEMMA 3.2. *If* $X = H_0^1(a, b)$, $Y = L^2(a, b)$, $F: X \to Y$ *is given in* (1.2), *and* $\partial k/\partial x(t, \tau, x)$ *is continuous in all its arguments, then* F *is weakly continuous.*

Proof. Suppose $\{x_n\}$ converges weakly to x in $H^1(a, b)$. Then there exists $\gamma > 0$ such that $\gamma \geqq \|x_n\|_\infty := \sup\{|x_n(\tau)|: a \leqq \tau \leqq b\}$, and by the mean value theorem,

$$\begin{aligned} |F(x_n)(t) - F(x)(t)| &\leqq \int_a^b \left|\frac{\partial k}{\partial \eta}(t, \tau, \eta(\tau))\right| \cdot |x_n(\tau) - x(\tau)|\,d\tau \\ &\leqq (b-a)C\|x_n - x\|_\infty, \end{aligned}$$

where $\eta(\tau)$ lies between $x_n(\tau)$ and $x(\tau)$ and

$$C = \max\{|\partial k/\partial \eta(t, \tau, \eta)|: a \leqq t, \tau < b, |\eta| \leqq \gamma\}.$$

Consequently,

$$\|F(x_n) - F(x)\| \leqq (b-a)^{3/2}C\|x_n - x\|_\infty.$$

The lemma follows from the fact that weak convergence in $H^1(a, b)$ implies strong convergence in $C[a, b]$. □

COROLLARY 3.3. *Let* $X = H_0^1(a, b)$, $Y = L^2(a, b)$. *For* $F: X \to Y$ *given in* (3.1) *and* β *sufficiently small, problem* (1.4) *has a solution.*

Proof. Let

$$k(t, \tau, x) = \log\left[\frac{(t-\tau)^2 + H^2}{(t-\tau)^2 + (H-x)^2}\right]. \tag{3.4}$$

Then

$$\frac{\partial k}{\partial x}(t, \tau, x) = \frac{2(H-x)}{(t-\tau)^2 + (H-x)^2}. \tag{3.5}$$

If β is sufficiently small, then $H - x(\tau) > 0$ for each $\tau \in [a, b]$, and $\partial k/\partial x$ is continuous. Thus by Lemma 3.2, F is weakly continuous, and by Theorem 2.1, (1.4) has a solution. □

Remark 3.4. In practice, problem (1.4) must be solved numerically. Let P_n denote a projection of X onto an n-dimensional subspace X_n, and suppose $P_n^T \to I_X$ (pointwise convergence to the identity in X). Similarly, let Q_m denote a projection of Y onto an m-dimensional subspace Y_m with $Q_m \to I_Y$ and suppose the Q_m's are uniformly bounded. Define

$$F_{mn}(x) := Q_m F(P_n x). \tag{3.6}$$

Since F is weakly continuous, each F_{mn} is weakly continuous and by Theorem 2.1 we can find a solution x_{mn} to each problem

$$\min_{x \in X} \|F_{mn}(x) - y_m\|^2 \quad \text{subject to } J(x) \leqq \beta_{mn}.$$

Suppose x_{mn} converges weakly to some x_*. For each $u \in X$,

$$\begin{aligned} |\langle P_n x_{mn} - x_*, u\rangle| &\leqq |\langle x_{mn}, P_n^T u - u\rangle| + |\langle x_{mn} - x_*, u\rangle| \\ &\leqq \|x_{mn}\| \, \|(P_n^T - I_X)u\| + |\langle x_{mn} - x, u\rangle|, \end{aligned}$$

so $P_n x_{mn}$ converges weakly to x_*. Then by weak continuity of F,

$$\|F_{mn}(x_{mn}) - F(x_*)\| \leqq \|Q_m\| \, \|F(P_n x_{mn}) - F(x_*)\| + \|(Q_m - I)F(x_*)\| \to 0.$$

This shows F_{mn} converges to F in the sense of assumption (A2) in § 2.

Remark 3.5. The operator F in (3.1) is twice (Frechet) differentiable. The first derivative $F'(x)$: $X \to Y$ is given by

$$[F'(x)u](t) = \int_a^b \frac{\partial k}{\partial x}(t, \tau, x(\tau))u(\tau)\, d\tau, \qquad a \leqq t \leqq b,\ x, u \in X, \tag{3.7}$$

where the kernel $\partial k/\partial x(t, \tau, x)$ is defined in (3.5). $F'(x)$ is compact, since its kernel is square integrable. The second derivative $F''(x): X \times X \to Y$ is given by

$$[F''(x)(u, v)](t) = \int_a^b \frac{\partial^2 k}{\partial x^2}(t, \tau, x(\tau))u(\tau)v(\tau)\, d\tau, \qquad a \leqq t \leqq b,\ x, u, v \in X. \tag{3.8}$$

4. Numerical solution of the regularized problem. To numerically solve (1.4), we obtain a finite-dimensional problem by choosing linearly independent basis functions $\{\phi_j\}_{j=1}^n \subset X = H_0^1(a, b)$ and taking approximations

$$\tilde{x}(\tau) = \sum_{j=1}^n x_j \phi_j(\tau), \qquad x = [x_1, \cdots, x_n]^T \in R^n. \tag{4.1}$$

We also approximate the norm in $Y = L^2(a, b)$ by a discrete sum

$$\|y\| \approx \sqrt{\frac{1}{m} \sum_{i=1}^{m} y(t_i)^2}, \qquad \{t_i\}_{i=1}^{m} \subset [a, b]. \tag{4.2}$$

To put this approximation in the context of the projection operator Q_m of Remark 3.4, we might define

$$Q_m y = \sum_{i=1}^{m} \langle \psi_i, y \rangle \psi_i = \sum_{i=1}^{m} \int_a^b \psi_i(t) y(t)\, dt\, \psi_i,$$

where each ψ_i is a nonnegative "averaging function" whose support consists of disjoint subintervals of $[a, b]$. By the smoothness of $y = F(x)$ (see Remark 3.1), we can apply the mean value theorem for integrals to obtain

$$\langle \psi_i, y \rangle = y(t_i) \int_a^b \psi_i(t)\, dt$$

for some point t_i in the ith subinterval. By scaling the ψ_i's appropriately and by their orthogonality, we obtain the right-hand side of (4.2) by taking $\|Q_m y\|$.

Thus the integral operator F in (3.1) gives rise to an operator $F_{mn}: R^n \to R^m$:

$$[F_{mn}(x)]_i := \int_a^b k(t_i, \tau, \tilde{x}(\tau)), \qquad 1 \leqq i \leqq m. \tag{4.3}$$

Similarly the derivative operator $F'(x)$ in (3.5), (3.7) yields an $m \times n$ matrix:

$$[F'_{mn}(x)]_{ij} = \int_a^b \frac{\partial k}{\partial x}(t_i, \tau, \tilde{x}(\tau)) \phi_j(\tau)\, d\tau, \qquad 1 \leqq i \leqq m, 1 \leqq j \leqq n. \tag{4.4}$$

From (2.1) and (4.1), the constraint $J(\tilde{x}) \leqq \beta^2$ yields $x^T B x \leqq \beta^2$, where B is the symmetric, positive definite matrix with entries

$$[B]_{jk} = \langle B\phi_j, \phi_k \rangle = \int_a^b \frac{d}{d\tau}[(B\phi_j)(\tau)]\, d\tau, \qquad 1 \leqq j, k \leqq n. \tag{4.5}$$

Without loss of generality we may assume $B = I$. Otherwise, we can compute a Choleski factorization $B = R^T R$, R nonsingular, and consider the change of variables $\hat{x} = Rx$. Hence we take as our (finite-dimensional) penalty functional

$$J(x) := \|x\|^2 = x^T x.$$

Remark 4.1. If B is not strictly positive definite, one may apply a similar change of variables and consider the penalty functional $J(x) = \|Px\|^2$, where P is the orthogonal projection onto the orthogonal complement of the null space of B. Computational details appear in the paper by Elden [5].

The resulting finite-dimensional analogue of problem (1.4) is then

$$\min_{x \in \mathbf{R}^n} \|F_{mn}(x) - y\|^2 \quad \text{subject to } J(x) := \|x\|^2 \leqq \beta^2. \tag{4.6}$$

We implemented several constrained optimization codes from the widely available Numerical Algorithms Group (NAG) software library to solve (4.6). We found both the Augmented Lagrangian code and the Sequential Quadratic Programming code in the NAG library to be unreliable for our model problem (3.1) for moderately large values of the parameter H (e.g., $H \approx 0.1$). We suspect that this lack of robustness is due to deficiencies in the line search phase of these algorithms. The line searches rely

on merit functions which try to balance the often conflicting requirements of reducing the objective function and maintaining the constraint. Solutions to highly ill-conditioned problems appear to be very sensitive to parameters which determine this balance.

Trust region methods have long been popular for *unconstrained* optimization problems

$$\min_x f(x) \tag{4.7}$$

where $f: R^n \to R$ is "smooth." See Dennis and Schnabel [4] for a discussion of convergence theory and numerical implementation. In their simplest form, trust region methods generate a new approximation $x_{k+1} = x_k + s$ from the current approximation x_k as follows: One takes a quadratic approximation $Q(s)$ to $f(x_k + s)$ and then solves the subproblem

$$\min Q(s) \quad \text{subject to } \|s\|^2 \leqq \delta_k^2, \tag{4.8}$$

where $\delta_k > 0$. If the solution s decreases the objective functional, i.e., if $f(x_k + s) < f(x_k)$, one sets $x_{k+1} = x_k + s$ and proceeds. Otherwise, one decreases the trust region parameter δ_k and resolves (4.8) until either $f(x_k + s) < f(x_k)$ or $\delta_k \approx 0$, in which case the iteration is terminated.

To robustly solve the *constrained* problem (4.6), we consider the trust region iteration

$$x_{k+1} = x_k + s_k, \qquad k = 0, 1, \cdots,$$

where s_k solves the quadratic subproblem

$$\min_s Q(s) := \|As - b\|^2, \quad \text{where } A := F'_{mn}(x_k), \quad b := y - F_{mn}(x_k), \tag{4.9}$$

subject to

$$\begin{bmatrix} J(x_k + s) - \beta^2 \\ J(s) - \delta_k^2 \end{bmatrix} = \begin{bmatrix} (x_k + s)^T (x_k + s) - \beta^2 \\ s^T s - \delta_k^2 \end{bmatrix} \leqq 0. \tag{4.10}$$

At each iteration, the trust region parameter $\delta_k > 0$ is chosen so the objective function

$$f(x) := \|F_{mn}(x) - y\|^2 \tag{4.11}$$

is reduced. Note that Q in (4.9) is the usual Gauss-Newton approximation to f, which is obtained from the Taylor expansion $F_{mn}(x_k + s) = F_{mn}(x_k) + F'_{mn}(x_k)s + O(\|s\|^2)$.

The constraint region in (4.10) is convex and is nonempty provided $J(x_k) \leqq \beta^2$. If $A := F'_{mn}(x_k)$ has full column rank, then $Q(s)$ is strictly convex, and subproblem (4.9), (4.10) has a unique solution. The following theorem shows that when δ_k is small, our trust region method behaves like a projected gradient method. We will use "∇" to indicate derivative with respect to x.

THEOREM 4.2. *Suppose* $\beta^2 > 0$, $x_k \neq 0$, $J(x_k) \leqq \beta^2$, *and* $\nabla f(x_k) \neq 0$, *but* x_k *does not satisfy the first-order necessary conditions* (2.2)-(2.4). *Let* s *solve the quadratic subproblem* (4.9), (4.10). *Then*

$$\frac{s}{\|s\|} \to \frac{d}{\|d\|} \quad as \quad \delta_k \to 0,$$

where

$$d = -\nabla f(x_k), \quad if \quad \nabla f(x_k)^T \nabla J(x_k) > 0 \quad or \; if \quad J(x_k) < \beta^2. \tag{4.12}$$

Otherwise, $\nabla f(x_k)^T \nabla J(x_k) \leqq 0$, $J(x_k) = \beta^2$, *and*

$$d = -\nabla f(x_k) + \frac{\nabla f(x_k)^T \nabla J(x_k)}{\|\nabla J(x_k)\|^2} \nabla J(x_k). \tag{4.13}$$

Note that (4.12) *gives the negative gradient, or "steepest descent" direction for the objective function f. Equation* (4.13) *gives the projected gradient direction, which is the orthogonal projection of* $-\nabla f(x_k)$ *onto the tangent subspace* $\{s \in X: \nabla J(x_k)^T s = 0\}$.

Proof. To simplify notation, define $\nabla f := \nabla f(x_k) = -2A^T b$ and $\nabla J := \nabla J(x_k) = 2x_k$. First-order necessary conditions for a solution to (4.9), (4.10) give

$$\begin{aligned} s &= [A^T A + (\lambda + \mu) I]^{-1} (A^T b - \lambda x_k) \\ &= \tfrac{1}{2} [A^T A + (\lambda + \mu) I]^{-1} (-\nabla f - \lambda \nabla J), \end{aligned}$$

where $\lambda \geqq 0$, $\mu \geqq 0$, $J(x_k + s) \leqq \beta^2$, $J(s) \leqq \delta_k^2$, and $\lambda[J(x_k + s) - \beta^2] + \mu[J(s) - \delta_k^2] = 0$. Note that $\delta_k \to 0 \Leftrightarrow J(s) = \|s\|^2 \to 0 \Leftrightarrow \mu \to \infty$. Similarly, a straightforward calculation shows that

$$x_k + s = [A^T A + (\lambda + \mu) I]^{-1} [A^T (A x_k + b) + \mu x_k],$$

so that $\lambda \to \infty \Leftrightarrow J(x_k + s) = \|x_k + s\|^2 \to 0$. But then the constraint $J(x_k + s) \leqq \beta^2$ would become inactive, and $\lambda = 0$. This contradiction shows that λ must remain bounded. Thus for small δ^k,

$$s = \frac{-\nabla f - \lambda \nabla J}{2\mu} + O\left(\frac{1}{\mu^2}\right). \tag{4.14}$$

Since $\nabla f = -2A^T b$, the objective functional for the quadratic subproblem (4.9), (4.10) can then be expressed as

$$Q(s) = \frac{-\|\nabla f\|^2 - \lambda \nabla f^T \nabla J}{\mu} + O\left(\frac{1}{\mu^2}\right) + \|b\|^2. \tag{4.15}$$

Similarly, $\nabla J = 2x_k$ and $J(x_k + s) = J(x_k) + 2x_k^T s + J(s)$ yields

$$J(x_k + s) - \beta^2 = \frac{-\nabla f^T \nabla J - \lambda \|\nabla J\|^2}{\mu} + O\left(\frac{1}{\mu^2}\right) + J(x_k) - \beta^2 \leqq 0. \tag{4.16}$$

If $\nabla f^T \nabla J > 0$ or $J(x_k) < \beta^2$, we see from (4.16) that the constraint becomes inactive for μ sufficiently large, in which case the complementarity condition forces $\lambda = 0$. We then obtain (4.12) from (4.14) as $\delta_k \to 0$. On the other hand, if $\nabla f^T \nabla J \leqq 0$ and $J(x_k) = \beta^2$, (4.16) gives

$$\lambda \geqq \frac{-\nabla f^T \nabla J}{\|\nabla J\|^2} + O\left(\frac{1}{\mu}\right). \tag{4.17}$$

To minimize (4.15), we take equality in (4.17). Then as $\delta_k \to 0$, $\lambda \to -(\nabla f^T \nabla J / \|\nabla J\|^2)$, and we obtain (4.13) from (4.14). □

The following corollary shows that for δ_k sufficiently small, the objective function f is decreased.

COROLLARY 4.3. *Suppose the conditions of Theorem* 4.2 *hold. Then for s the solution to* (4.9), (4.10) *and* δ_k *sufficiently small,* $f(x_k + s) < f(x_k)$.

Proof. Since f is twice continuously differentiable,

$$f(x_k + s) = f(x_k) + \nabla f(x_k) s + O(\|s\|^2).$$

Hence, it suffices to show that there exists $\gamma > 0$ for which

$$\nabla f(x_k) s \leqq -\gamma \|s\| \tag{4.18}$$

whenever δ_k is sufficiently small. By Theorem 4.2 we have that as $\delta_k \to 0$, either

$$\nabla f(x_k)^T \frac{s}{\|s\|} \to -\|\nabla f(x_k)\|^2,$$

or else

$$\nabla f(x_k)^T \frac{s}{\|s\|} \to -\|\nabla f(x_k)\|^2 + \frac{|\nabla f(x_k)^T \nabla J(x_k)|^2}{\|\nabla J(x_k)\|^2}.$$

In this second case, by Schwartz's inequality,

$$|\nabla f(x_k)^T \nabla J(x_k)| \leqq \|\nabla f(x_k)\| \, \|\nabla J(x_k)\|,$$

with equality if and only if $\nabla f(x_k) = \lambda \nabla J(x_k)$. Since we have assumed the first-order necessary conditions (2.2)-(2.4) do not hold, Schwartz's inequality is strict. In either case,

$$\lim_{\delta_k \to 0} \nabla f(x_k)^T \frac{s}{\|s\|} < 0.$$

Equation (4.18) follows from the continuity of the solution s with respect to δ_k. □

This trust region approach gives decrease in the objective function outside the region of convergence of the Gauss-Newton method. Once we are inside this region of convergence, we take the Gauss-Newton step, the trust region constraint $J(s) \leqq \delta_k^2$ becomes inactive, and we obtain the following result. Note that

$$\nabla^2 f(x) = 2[F'(x)^T F'(x) + N(x)],$$

where

$$N(x) := F''(x)^T (F(x) - y).$$

THEOREM 4.4. *Define* $c(x) := J(x) - \beta^2 = \|x\|^2 - \beta^2$, *let* x_* *be a solution to* (4.6) *with corresponding Lagrange multiplier* λ_*, *and suppose the second-order sufficient condition*

$$s^T [\nabla^2 f(x_*) + \lambda_* \nabla^2 c(x_*)] s \geqq \alpha \|s\|^2 \quad \text{whenever } \lambda_* \nabla c(x_*)^T s = 0 \tag{4.19}$$

holds for some $\alpha > 0$. *Assume each* δ_k *has been chosen so the trust region constraint* $J(s) \leqq \delta_k$ *is inactive. If* $N(x_*)$ *is sufficiently small and* x_0 *is sufficiently close to* x_*, *then iteration* (4.9), (4.10) *converges to* x_*. *If* $N(x_*) = 0$, *the rate of convergence is locally q-quadratic. Otherwise, the rate is linear.*

Proof. If $\lambda_* = 0$, the standard analysis for the unconstrained Gauss-Newton method applies. See for example, Theorem 10.2.1 and Corollary 10.2.2 of Dennis and Schnabel [4]. Otherwise, x_* and $\lambda_* > 0$ are locally unique solutions to

$$\begin{aligned} \nabla f(x) + \lambda \nabla c(x) &= 0, \\ c(x) &= 0. \end{aligned} \tag{4.20}$$

For x_k and $x_{k+1} = x_k + s$ sufficiently close to x_* and s the solution to (4.9), (4.10), $c(x_k) = c(x_k + s) = 0$, and first-order necessary conditions give

$$\begin{aligned} \nabla Q(s) + \lambda_{k+1} \nabla c(x_k + s) &= 0, \\ c(x_k + s) &= 0. \end{aligned}$$

Since $\nabla Q(s) = \nabla f(x_k) + 2A^T As$ and $0 = c(x_k) = c(x_k + s) = c(x_k) + \nabla c(x_k)^T s + \frac{1}{2} s^T \nabla^2 c s$, where $\nabla^2 c = 2I$, setting $\lambda_{k+1} = \lambda_k + \Delta\lambda$ above gives

$$\begin{bmatrix} 2A^TA + \lambda_k \nabla^2 c & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} s \\ \Delta\lambda \end{bmatrix} = \begin{bmatrix} \nabla f(x_k) + \lambda_k \nabla c(x_k) \\ c(x_k) \end{bmatrix} + O(\Delta\lambda \|s\| + \|s\|^2).$$

The theorem follows from standard analysis of quasi-Newton methods for solving (4.20). As in the unconstrained case, if $N(x_*) = 0$, we obtain local quadratic convergence. If $0 < (\|N(x_*)\|/\alpha) < 1$, where α is the coercivity constant in (4.19), we obtain linear convergence. □

Remark 4.5. The quantity $(\|N(x_*)\|/\alpha)$ in the above proof governs the rate of local (linear) convergence when $N(x_*) \neq 0$. By an argument similar to that in Remark 2.7, we see that for large values of λ (which correspond to small values of the regularization parameter β) α is large. As λ decreases, one would expect α to also decrease. In this case, one observes much slower convergence of iteration (4.9), (4.10).

Solution to the quadratic subproblem. To solve subproblem (4.9), (4.10) in a numerically stable manner and to reduce the computational cost, we first diagonalize it using an approach similar to that of Elden [5]. We will also use the diagonal entries to determine a reasonable choice for the regularization parameter in § 5. We assume $A = F'(x_k)$ has full column rank. Let A have the singular value decomposition

$$A = UDV^T,$$

where U and V are orthogonal matrices and $D = \operatorname{diag}\{d_i\}$ has the (positive) singular values d_i of A as its diagonal entries. Subproblem (4.9), (4.10) is then equivalent to

$$\min_{\hat{s}} \|D\hat{s} - \hat{b}\|^2 \tag{4.21}$$

subject to

$$\begin{bmatrix} \|\hat{x} + \hat{s}\|^2 - \beta^2 \\ \|\hat{s}\|^2 - \delta_k^2 \end{bmatrix} \leqq 0, \tag{4.22}$$

where

$$\hat{s} := V^T s, \quad \hat{x} := V^T x, \quad \hat{b} := U^T b.$$

Our approach to solving (4.21), (4.22) follows ideas outlined in Pang's paper [12, p. 65]. First-order necessary conditions for the solution $\hat{s}$ are

$$\begin{gathered} D^T[D\hat{s} - \hat{b}] + \lambda(\hat{x} + \hat{s}) + \mu\hat{s} = 0, \\ \|\hat{x} + \hat{s}\|^2 \leqq \beta^2, \qquad \|\hat{s}\|^2 \leqq \delta_k^2, \\ \lambda \geqq 0, \qquad \mu \geqq 0, \\ \lambda[\|\hat{x} + \hat{s}\|^2 - \beta^2] + \mu[\|\hat{s}\|^2 - \delta_k^2] = 0. \end{gathered} \tag{4.23}$$

From this we obtain

$$\hat{s}(\bar{\lambda}) = [D^TD + (\lambda + \mu)I]^{-1}(D^T\hat{b} - \lambda\hat{x}), \tag{4.24}$$

$$\bar{\lambda} := \begin{bmatrix} \lambda \\ \mu \end{bmatrix} \geqq 0, \quad \bar{c}(\bar{\lambda}) := \begin{bmatrix} \|\hat{x} + \hat{s}(\bar{\lambda})\|^2 - \beta^2 \\ \|\hat{s}(\bar{\lambda})\|^2 - \delta_k^2 \end{bmatrix} \leqq 0, \qquad \bar{\lambda}^T \bar{c}(\bar{\lambda}) = 0. \tag{4.25}$$

Thus the Lagrange multipliers λ, μ are solutions to the nonlinear (quadratic) complementarity problem (4.25). Once these have been obtained, $\hat{s}$ is computed from (4.24), and $s = V\hat{s}$ is the solution to subproblem (4.9), (4.10).

We solve the quadratic complementarity problem (4.25) using the following Newton iteration: For $k=0, 1, \cdots$, let $\bar{\lambda}=\bar{\lambda}_{k+1}$ solve the linear complementarity problem

$$\bar{\lambda} \geqq 0, \tag{4.26}$$

$$\bar{c}(\bar{\lambda}_k)+\nabla \bar{c}(\bar{\lambda}_k)(\bar{\lambda}-\bar{\lambda}_k) \leqq 0, \tag{4.27}$$

$$\bar{\lambda}^T[\bar{c}(\bar{\lambda}_k)+\nabla \bar{c}(\bar{\lambda}_k)(\bar{\lambda}-\bar{\lambda}_k)]=0. \tag{4.28}$$

For a summary of convergence results of iteration (4.26)-(4.28) as well as a review of relevant literature, we refer the reader to Pang [12]. Note that this iteration is locally quadratically convergent. To obtain global convergence to a solution of (4.25), we added a line search. Define

$$h(\bar{\lambda}) := \min(\bar{\lambda}, \bar{c}(\bar{\lambda})),$$

(componentwise minimum) and observe that $\bar{\lambda}$ solves (4.25) if and only if $h(\bar{\lambda})=0$. Given a solution $\bar{\lambda}$ to (4.26)-(4.28) for which $\|h(\bar{\lambda})\|_1 \geqq \|h(\bar{\lambda}_k)\|_1$, we define the Newton step

$$\Delta=\bar{\lambda}-\bar{\lambda}_k,$$

and obtain a solution γ_* to the one-dimensional minimization problem

$$\min_{0 \leqq \gamma < 1} \|h(\bar{\lambda}_k+\gamma\Delta)\|_1.$$

We then take

$$\bar{\lambda}_{k+1}=\bar{\lambda}_k+\gamma_*\Delta$$

as the new estimate for the solution to (4.25).

5. Numerical results. We applied our constrained least squares regularization method to the model problem of § 3. All computations were performed on a Zenith (IBM-compatible) AT Personal Computer. F is the nonlinear integral operator in (3.1) with $a=0$, $b=1$, and $H=0.2$. We took approximate solutions x from the n-dimensional subspace of $H_0^1(0,1)$ spanned by piecewise linear functions

$$\phi_j(t) := \begin{cases} \dfrac{\tau-\tau_j}{h}, & \text{if } \tau_{j-1} \leqq \tau \leqq \tau_j, \\ \dfrac{\tau_{j+1}-\tau}{h} & \text{if } \tau_j \leqq \tau \leqq \tau_{j+1}, \\ 0, & \text{otherwise,} \end{cases} \tag{5.1}$$

where $\tau_j=jh$, $h=1/(n+1)$, $j=1, \cdots, n=25$. The integrals (4.3), (4.4) were computed numerically. We took as our true solution a linear combination of two Gaussians,

$$x_{\text{true}}(\tau)=c_1 \exp(d_1(\tau+p_1)^2)+c_2 \exp(d_2(\tau-p_2)^2)+c_3\tau+c_4,$$

where $c_1=-0.1$, $c_2=-0.075$, $d_1=-40$, $d_2=-60$, $p_1=0.4$, $p_2=0.67$, and c_3, c_4 are chosen so that $x(0)=x(1)=0$. $H-x_{\text{true}}(\tau)$ is plotted in Fig. 1. We took data points $y_i=y(t_i)+\varepsilon_i$, $t_i=i/(m+1)$, $i=1, \cdots, m=30$. The ε_i simulate measurement errors and are pseudorandom and normally distributed with mean 0 and variance σ^2 chosen so that the noise to signal ration was 0.2 percent, i.e.,

$$\frac{\sigma}{\|F_{mn}(x)\|}=.002.$$

We took as our (infinite-dimensional) penalty functional

$$J(x) = \|x\|^2 = \int_0^1 x'(\tau)^2 \, d\tau.$$

Figure 2 is a semilog plot of the singular values of the derivative $A = F'_{mn}(x)$ at the initial guess $x_0(\tau) \equiv 0$. The exponential decay rate of the singular values gives a quantitative indication of the server ill-posedness of the problem.

We solved a sequence of finite-dimensional problems (4.6) with increasing $\beta \in \{.2, .25, .275, .3, .35, .4, .5,\}$ using the trust region algorithm of the previous section. Resulting approximate solutions x_β are shown in Fig. 3 for $\beta = .2$ (dashed line), $\beta = .275$ (solid line), and $\beta = .5$ (dotted line). The true solution satisfies $J(x_{\text{true}}) = (.277)^2$ and is represented by +'s. The constraint $J(x) \leqq \beta^2$ was active in each case.

Figure 4 is a semilog plot of the corresponding Lagrange multipliers $\lambda = \lambda(\beta)$. From Remark 2.7, we would expect the regularization problem (4.6) to become highly ill-conditioned for larger values of β. Also, by Remark 4.5, we would expect the rate of convergence of iteration (4.9), (4.10) to slow considerably in the presence of noisy data. We have observed both these phenomena in our numerical experiments.

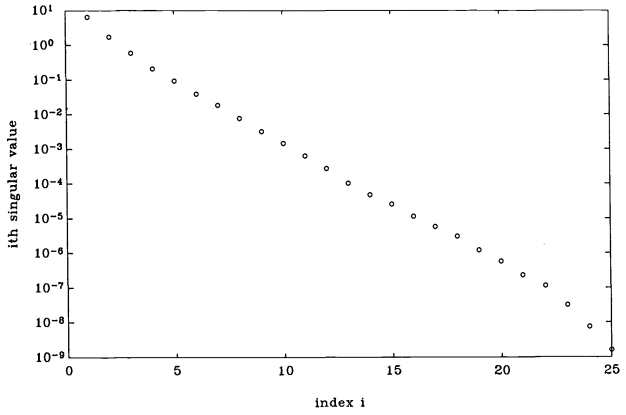


FIG. 2. *Semilog plot of singular values of discrete derivative operator in decreasing order of magnitude.*

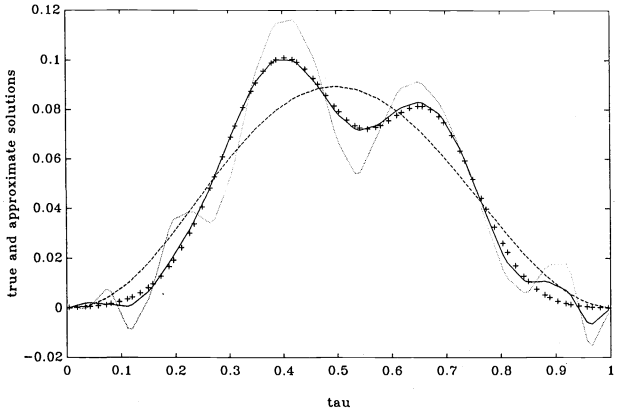


FIG. 3. *Approximate solutions* $-x_\beta$, *for* $\beta = .2$ (*dashed line*), $\beta = .275$ (*solid line*), *and* $\beta = .5$ (*dotted line*). *The* +*'s represent the negative true solution,* $-x_{\text{true}}$.

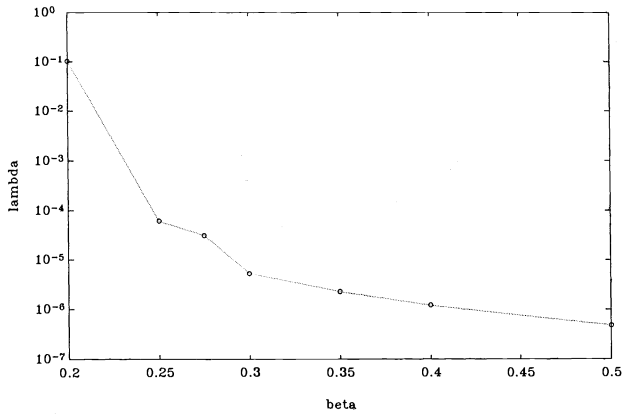


FIG. 4. *Semilog plot of the Lagrange multipliers λ versus the regularization parameter β.*

Figure 5 gives the error indicators as a function of the regularization parameter β. The curves represent, from top to bottom, (i) the "solution error"

$$e(\beta) := J(x_\beta - x_{\text{true}}) = \|x_\beta - x\|^2; \tag{5.2}$$

(ii) the generalized cross validation functional $V(\beta)$, defined in (5.5) below; and (iii) the weighted objective functional obtained from (4.11) and (4.3),

$$\frac{1}{m} f(x_\beta) := \frac{1}{m} \sum_{i=1}^{m} [F_{mn}(x_\beta)(t_i) - y_i]^2. \tag{5.3}$$

The fact that the solution error $e(\beta)$ increases for large β while the objective functional $f(x_\beta)$ continues to decrease is another consequence of the ill-posedness of this problem.

Remark 5.1. The practical choice of a regularization parameter for a given error-contaminated data set is a difficult matter. Many methods require prior knowledge of the magnitude of the error and/or the norm of the true solution. A statistical technique known as the method of Generalized Cross Validation (GCV) requires only that the error be random in the sense that

$$E(\varepsilon_i) = 0; \qquad E(\varepsilon_i \varepsilon_j) = \sigma^2 \delta_{ij}. \tag{5.4}$$

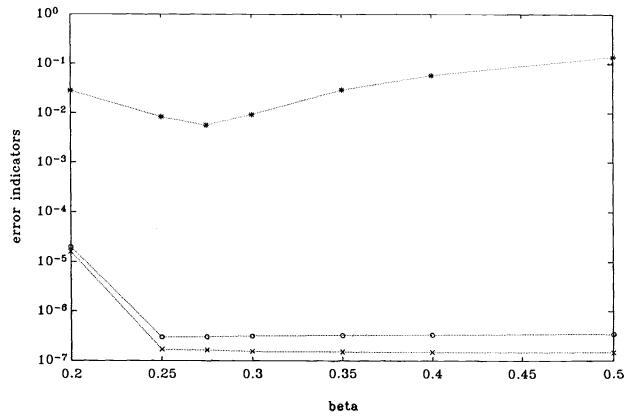


FIG. 5. *Semilog plot of error indicators: Solution error $e(\beta)$ (represented by *'s); GCV function $V(\beta)$ (represented by o's); and scaled objective functional $f(x_\beta)/m$ (represented by x's).*

GCV has been successfully applied to a variety of regularization methods including Tikhonov Regularization (see [15], [10]) and Truncated Singular Value Decomposition (See [14]). For the regularization problem (4.6), the GCV functional is given by

$$V(\beta) = \frac{\frac{1}{m} f(x_\beta)}{\left\{\frac{1}{m}\,\text{Trace}\,[I_m - A(A^TA+\lambda I)^{-1}A^T]\right\}^2}, \tag{5.5}$$

$$= \frac{\frac{1}{m} f(x_\beta)}{\left\{\frac{1}{m}\left[(m-n)+\lambda \sum_{i=1}^{n} \frac{1}{d_i^2+\lambda}\right]\right\}^2}$$

where $\lambda = \lambda(\beta)$ is the Lagrange multiplier for problem (4.6) and the d_i's are the singular values of the derivative $A = F'_{mn}(x_\beta)$. In Fig. 5 we see that although V is very flat, the minimum of $V(\beta)$ coincides with the minimum of the solution error $e(\beta)$. At least for this particular example, the minimizer of $V(\beta)$ provides a very reasonable choice for the regularization parameter.

REFERENCES

[1] P. DEUFLHARD AND F. HAIRER, EDS. (1983), *Numerical Treatment of Inverse Problems in Differential and Integral Equations*, Birkhauser, Boston.

[2] W. ALT (1983), *Lipschitzian perturbations of infinite dimensional problems*, in Mathematical Programming with Data Perturbations II, A. V. Fiaco, ed., Lecture Notes in Pure and Applied Mathematics 85, Marcel Dekker, New York.

[3] F. COLONIUS AND K. KUNISCH (1986), *Output least squares is elliptic systems*, Tech. Report 86-76, Institute fur Mathematik, Technische Universitat Graz, Kopernikusgasse 24, A-8010, Graz, Austria.

[4] J. E. DENNIS AND R. B. SCHNABEL (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall, Englewood Cliffs, NJ.

[5] L. ELDEN (1977), Algorithms for the regularization of ill-conditioned least squares problems, BIT 17, pp. 134-145.

[6] C. KRAVARIS AND J. H. SEINFELD (1985), *Identification of parameters in distributed parameter systems by regularization*, SIAM J. Control Optim., 23, pp. 217-241.

[7] G. KRISTENSSON AND C. R. VOGEL (1986), *Inverse, scattering for acoustic waves using the penalized likelihood method*, Inverse Problems, 2, pp. 461-479.

[8] D. G. LUENBERGER (1969), *Optimization by Vector Space Methods*, John Wiley, New York.

[9] A. I. PRILEPKO (1965), *Uniqueness of determination of the form of a body from the values of the external potential*, Doklady Akad. Nauk SSSR, 160.

[10] F. O'SULLIVAN AND G. WAHBA (1985), *A cross validated Bayesian retrieval algorithm for nonlinear remote sensing experiments*, J. Comput. Phys., 59, pp. 551-555.

[11] T. I. SEIDMAN AND C. R. VOGEL (1987), *Well-posedness and convergence of some regularization methods for nonlinear ill-posed problems*, Tech. Report No. CMA-R48-86, Centre for Mathematical Analysis, The Australian National University, Canberra, Australia.

[12] J. S. PANG (1986), *Inexact Newton methods for the nonlinear complementarity problem*, Math. Programming, 36, pp. 54-71.

[13] A. N. TIKHONOV AND V. Y. ARSENIN (1977), *Solutions of Ill-Posed Problems*, John Wiley, New York.

[14] C. R. VOGEL (1986), *Optimal choice of a truncation level for the truncated* SVD *solution of linear first kind integral equations when data are noisy*, SIAM J. Numer. Anal., 23, pp. 109-117.

[15] G. WAHBA (1977), *Practical approximate solutions to linear operator equations when the data are noisy*, SIAM J. Numer. Anal., 14, pp. 651-667.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 50-69, January 1990

003

THE SENSITIVITY OF THE ALGEBRAIC AND DIFFERENTIAL RICCATI EQUATIONS*

CHARLES KENNEY† AND GARY HEWER‡

Abstract. In this paper it is shown that the ideas developed by Byers in [*Proc. Summer Research Conference*, AMS Vol. 47, Contemporary Math., American Mathematical Society, Providence, RI, 1984, pp. 35-49] on the sensitivity of the algebraic Riccati equation can be sharpened and extended to norms other than the Frobenius norm. This extension is crucial from an interpretive point of view because use of the spectral norm allows an identification between the condition number of the algebraic Riccati equation and the damping properties of the closed-loop dynamical system. Moreover, this approach has the pleasant feature that it carries over to a completely parallel theory for the sensitivity of the differential Riccati equation, an area that has not been considered previously.

Key words. Riccati equation, condition number, closed-loop damping

AMS(MOS) subject classifications. 65F35, 65F30, 15A12

1. Introduction.

Algebraic and differential Riccati equations arise in the problem of optimal control of linear time-invariant systems of the form

$$\begin{aligned} &\dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = x_0, \\ &y(t) = Cx(t). \end{aligned} \tag{1.1}$$

Here $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{k\times n}$; x, u, and y are the state, input, and output vectors, respectively. For a given symmetric positive semidefinite matrix P_1 and a terminal time t_1, the goal of optimal control is to find the input $u = u(t)$ that minimizes the cost functional

$$J(u, P_1, t_1) \equiv \int_0^{t_1} x^T(t)C^TCx(t) + u^T(t)u(t)\,dt + x^T(t_1)P_1x(t_1). \tag{1.2}$$

In this case, the input function, u_{t_1}, which minimizes the cost functional, is [16]

$$u_{t_1}(t) = -B^TP(t)x(t) \quad \text{for } 0 \leqq t \leqq t_1. \tag{1.3}$$

In (1.3), P is the solution to the differential Riccati equation

$$\dot{P}(t) = -G - A^TP(t) - P(t)A + P(t)FP(t), \qquad P(t_1) = P_1 \tag{1.4}$$

where $F = BB^T$ and $G = C^TC$. If u_{t_1} in (1.3) is used in (1.1), we obtain the closed-loop system

$$\dot{x}(t) = (A - FP(t))x(t), \quad x(0) = x_0, \quad 0 \leqq t \leqq t_1. \tag{1.5}$$

A related procedure involves letting t_1 go to infinity in (1.2). If (G, A) is detectable and (A, F) is stabilizable, then the algebraic Riccati equation

$$0 = G + A^TX + XA - XFX \tag{1.6}$$

* Received by the editors January 21, 1987; accepted for publication (in revised form) April 24, 1989.
† Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106. This research was supported by the National Science Foundation and the Air Force Office of Scientific Research under grant ECS-87-18897, and the National Science Foundation under grant DMS88-00817.
‡ RF Missile Systems Branch, China Lake, California 93555.

has a unique symmetric positive semidefinite solution, X [4]. Moreover, as $t_1 \to \infty$ in (1.2), the initial matrix $P(0)$ in (1.4) converges to X, independently of the terminal matrix P_1, so long as $P_1 = P_1^T \geqq 0$. The resulting closed-loop system,

$$\dot{x}(t) = (A - FX)x(t), \qquad x(0) = x_0, \tag{1.7}$$

is stable in the sense that the eigenvalues of $A - FX$ have negative real part.

This paper has two objectives. The first is to study the sensitivity of the solutions to (1.4) and (1.6) to perturbations in the coefficient matrices A, F, and G. Good upper and lower sensitivity bounds for the algebraic Riccati equation were obtained by Byers in [2]. We will show that these results can be sharpened and extended to the differential Riccati equation.

Our second objective is to point out a very strong relationship between the sensitivity of the Riccati equations (1.4), (1.6) and the damping of the closed-loop systems (1.5), (1.7). We define the damping D of the closed-loop system (1.7) as

$$D \equiv \max_{\|x_0\|=1} \|x\|_{L_2} = \max_{\|x_0\|=1} \left[\int_0^\infty \|x(t)\|^2 \, dt \right]^{1/2} \tag{1.8}$$

where $\|\cdot\|$ denotes the 2-norm, $\|v\|^2 = \Sigma v_i^2$. The maximum in (1.8) is taken over solutions to (1.7). A similar definition of damping applies to (1.5).

For the algebraic Riccati equation, the connection between sensitivity and damping relies on the identity

$$\begin{aligned} \int_0^\infty \|x(t)\|^2 \, dt &= \int_0^\infty x^T(t)x(t) \, dt \\ &= \int_0^\infty x_0^T \, e^{A_C^T t} \, e^{A_C t} x_0 \, dt \\ &= x_0^T \int_0^\infty e^{A_C^T t} \, e^{A_C t} \, dt \, x_0 \\ &\equiv x_0^T H x_0. \end{aligned} \tag{1.9}$$

In (1.9), A_C denotes the closed-loop matrix $A - FX$ and the matrix H satisfies $\Omega(H) = -I$ with

$$\Omega(Z) \equiv A_C^T Z + Z A_C. \tag{1.10}$$

This identity shows that damping is related to the closed-loop Lyapunov operator Ω. Moreover, we show that $D^2 = \|H\| = \|\Omega^{-1}\|$, for the induced operator 2-norm

$$\|\Omega^{-1}\| = \max_{M \neq 0} \frac{\|\Omega^{-1}(M)\|}{\|M\|}. \tag{1.11}$$

When coupled with a slightly more general definition of damping, these results clearly show the connection between the dynamical behavior of the closed-loop system and the sensitivity of the associated Riccati equation. This complements the results of [9], in which the sensitivity of the solution to the Lyapunov equation $A^T X + XA = -W$, for A stable, was shown to be related to the damping behavior of the dynamical system $\dot{x} = Ax$.

The outline of the paper is as follows. In § 2, we develop sensitivity and damping results for the algebraic Riccati equation. Section 3 extends these results to the differential Riccati equation. Section 4 is devoted to numerical tests of the bounds in §§ 2 and 3.

2. Sensitivity of the algebraic Riccati equation. We will assume throughout this paper that F and G are positive semidefinite matrices with (A, F) stabilizable and (G, A) detectable. Let $X = X(A, F, G)$ denote the unique positive semidefinite symmetric solution to (1.6). Our goal is to investigate the variation in X with respect to changes in A, F, and G. More precisely, let $\tilde{A}$, $\tilde{F}$, and $\tilde{G}$ be matrices that are near A, F, and G with respect to the matrix 2-norm. (Except for Theorem 2.1, all the results in this paper are for the two-norm.) Define $\Delta A = \tilde{A} - A$, $\Delta F = \tilde{F} - F$, and $\Delta G = \tilde{G} - G$. We assume that $\tilde{F}$ and $\tilde{G}$ are of the form $\tilde{B}\tilde{B}^T$ and $\tilde{C}^T\tilde{C}$ for some $\tilde{B}$ and $\tilde{C}$, so we will require $\tilde{F}$ and $\tilde{G}$ to be symmetric and positive semidefinite. For $\|\Delta A\|$, $\|\Delta F\|$, and $\|\Delta G\|$ sufficiently small, $(\tilde{A}, \tilde{F})$ is stabilizable and $(\tilde{G}, \tilde{A})$ is detectable; hence the perturbed solution $\tilde{X} = X(\tilde{A}, \tilde{F}, \tilde{G})$ is well defined. Let $\Delta X = \tilde{X} - X$.

To relate $\|\Delta X\|$ to $\|\Delta A\|$, $\|\Delta F\|$, and $\|\Delta G\|$, we adopt the condition theory of Rice [19]. For sufficiently small $\delta > 0$, define $K_\delta = K_\delta(A, F, G)$ by

$$K_\delta \equiv \sup \left\{ \frac{\|\Delta X\|}{\delta \|X\|} \;\middle|\; \|\Delta A\| \leq \delta \|A\|, \|\Delta F\| \leq \delta \|F\|, \|\Delta G\| \leq \delta \|G\|, \tilde{F} \text{ and } \tilde{G} \right. \text{ symmetric positive semidefinite} \left. \right\}. \tag{2.1}$$

Taking the limit as δ goes to zero, we obtain the asymptotic condition number

$$K \equiv \lim_{\delta \to 0^+} K_\delta. \tag{2.2}$$

It is worth mentioning that this limit exists in an extended sense because K_δ is nonincreasing as δ goes to zero. Moreover, this limit is finite as we will see below.

We can obtain bounds on K by substituting $\tilde{A}$, $\tilde{F}$, $\tilde{G}$, and $\tilde{X}$ into (1.6). After some rearrangement, we find

$$\begin{aligned} A_C^T \Delta X + \Delta X A_C = &-\Delta G - \Delta A^T X - X \Delta A + X \Delta F X \\ &- (\Delta A - \Delta F X)^T \Delta X - \Delta X (\Delta A - \Delta F X) + \Delta X (F + \Delta F) \Delta X \end{aligned} \tag{2.3}$$

where $A_C = A - FX$ is the closed-loop matrix. The left-hand side of (2.3) has the form $\Omega(\Delta X)$, where Ω is defined by (1.10). Since A_C is stable, Ω is invertible [17] and

$$\Omega^{-1}(Z) = -\int_0^\infty e^{A_C^T t} Z e^{A_C t} \, dt. \tag{2.4}$$

We may rewrite (2.3) as

$$\begin{aligned} \Delta X = &-\Omega^{-1}(\Delta G + \Delta A^T X + X \Delta A - X \Delta F X) \\ &- \Omega^{-1}((\Delta A - \Delta F X)^T \Delta X + \Delta X (\Delta A - \Delta F X) - \Delta X (F + \Delta F) \Delta X). \end{aligned} \tag{2.5}$$

The first term on the right-hand side of (2.5) determines the norm of ΔX for ΔA, ΔG, and ΔF small, and is in fact the Fréchet derivative of the mapping $(A, F, G) \to X$. For the purposes of estimation and interpretation, it is convenient to break up this term into the sum of three linear operators (using the notation of Byers in [2]):

$$-\Omega^{-1}(\Delta G + \Delta A^T X + X \Delta A - X \Delta F X) = -\Omega^{-1}(\Delta G) - \Theta(\Delta A) + \Pi(\Delta F) \tag{2.6}$$

where

$$\Theta(Z) \equiv \Omega^{-1}(Z^T X + X Z), \qquad \Pi(Z) \equiv \Omega^{-1}(XZX). \tag{2.7}$$

In general, the operator Θ determines the sensitivity of X with respect to uncertainty in the state matrix A (e.g., modeling errors in the open-loop dynamics). Similarly,

Π and Ω^{-1} determine the sensitivity of X with respect to uncertainty in F (actuator errors) and G (sensors errors), respectively.

It is worth emphasizing that throughout this paper we are concerned only with the effects of "small" perturbations, i.e., we are dealing with a first-order sensitivity analysis.

Associated with (2.6) is the Byers approximate condition number

$$K_B \equiv \frac{\|\Omega^{-1}\|\,\|G\| + \|\Theta\|\,\|A\| + \|\Pi\|\,\|F\|}{\|X\|}. \tag{2.8}$$

The motivation for using this condition number is that if δ is small and $\|\Delta A\| \leqq \delta\|A\|$, $\|\Delta G\| \leqq \delta\|G\|$, $\|\Delta F\| \leqq \delta\|F\|$, then

$$\frac{\|\Delta X\|}{\delta\|X\|} \cong \frac{\|-\Omega^{-1}(\Delta G) - \Theta(\Delta A) + \Pi(\Delta F)\|}{\delta\|X\|}$$

$$\leqq \frac{\|\Omega^{-1}\|\,\|\Delta G\| + \|\Theta\|\,\|\Delta A\| + \|\Pi\|\,\|\Delta F\|}{\delta X} \leqq K_B.$$

Using a result from Stewart [20], Byers [2] was able to show that, for the Frobenius norm, K_B is a very good approximation to K.

THEOREM 2.1 (Byers [2]). *Define K and K_B by* (2.2) *and* (2.8) *for the Frobenius norm. Then* $(1/9)K_B \leqq K \leqq 4K_B$.

Proof. See [2] for the proof. □

Remark. From (2.7), $\|\Theta\| \leqq 2\|\Omega^{-1}\|\,\|X\|$ and $\|\Pi\| \leqq \|\Omega^{-1}\|\,\|X\|^2$. This leads to the "conservative" Byers condition number K_{CB}:

$$K_B \leqq K_{CB} \equiv \frac{\|\Omega^{-1}\|}{\|X\|}(\|G\| + 2\|A\|\,\|X\| + \|F\|\,\|X\|^2). \tag{2.9}$$

However, this simpler condition number is generally too conservative and is often several orders of magnitude larger than K_B (see [2]).

The next theorem shows that, for the matrix 2-norm, K_B is within a factor of 3 of K.

THEOREM 2.2. *Define K and K_B by* (2.2) *and* (2.8) *for the matrix 2-norm. Then* $\frac{1}{3}K_B \leqq K \leqq K_B$.

Proof. See Appendix 1 for the proof. □

These theorems show that K_B is entirely satisfactory as an estimator of K, especially in view of the fact that for most numerical purposes we only need to estimate K to within a factor of 10. However, the question arises as to what K_B means in terms of the original problem. While a geometric interpretation of ill-conditioning is difficult for the Frobenius norm, as noted in [2], it is rather easy for the spectral norm: X is sensitive to perturbations if the closed-loop system (1.7) is poorly damped with respect to the weighting matrices I, X, or X^2.

We define the damping of (1.7) with respect to X^k, for $k = 0, 1, 2$ by

$$D_k \equiv \max_{\|x_0\|=1} \|x\|_{L_2, X^k} \equiv \max_{\|x_0\|=1} \left[\int_0^\infty x^T(t) X^k x(t)\, dt\right]^{1/2}. \tag{2.10}$$

We will show that $D_0^2 = \|\Omega^{-1}\|$, $D_2^2 = \|\Pi\|$, and $2D_1^2 \leqq \|\Theta\| \leqq D_0 D_2$; thus establishing the intimate relationship between sensitivity and damping.

As a step in this direction, let H_k be the solution to the closed-loop Lyapunov equations,

$$A_C^T H_k + H_k A_C = -X^k \tag{2.11}$$

for $k = 0, 1, 2$. Since A_C is stable [17], $H_k = \int_0^\infty e^{A_C^T t} X^k e^{A_C t}\, dt$.

LEMMA 2.3. *Let x satisfy the closed-loop differential equation* (1.7). *Then* $x_0^T H_k x_0 = \|x\|^2_{L_2, X^k}$ *and* $\|H_k\| = \max_{\|x_0\|=1} \|x\|^2_{L_2, X^k} = D_k^2$.

Proof. Since $x(t) = e^{A_C t} x_0$, $\|x\|^2_{L_2, X^k} = \int_0^\infty x_0^T e^{A_C^T t} X^k e^{A_C t} x_0\, dt = x_0^T H_k x_0$. If $\|x_0\| = 1$ then $x_0^T H_k x_0 \leqq \|H_k\|$, because H_k is symmetric and positive definite. Thus $D_k^2 \leqq \|H_k\|$. Moreover, if we let x_0 be a unit eigenvector of H_k corresponding to $\lambda_{\max}(H_k) = \|H_k\|$, then we obtain equality: $D_k^2 = \|H_k\|$. □

The connection between damping and sensitivity relies on the observation that

$$H_0 = -\Omega^{-1}(I), \quad 2H_1 = -\Theta(I), \quad H_2 = -\Pi(I). \tag{2.12}$$

THEOREM 2.4. *Define* Ω, Θ, Π, *and* H_k *by* (1.10), (2.7), *and* (2.11). *Then*

$$\|\Omega^{-1}\| = \|H_0\|, \tag{2.13}$$

$$\|\Pi\| = \|H_2\|, \tag{2.14}$$

and

$$2\|H_1\| \leqq \|\Theta\| \leqq 2\|H_0\|^{1/2} \|H_2\|^{1/2}. \tag{2.15}$$

Proof. See Appendix 1 for the proof. □

COROLLARY 2.5.

$$D_0^2 = \|\Omega^{-1}\|, \quad D_2^2 = \|\Pi\|, \quad 2D_1^2 \leqq \|\Theta\| \leqq D_0 D_2. \tag{2.16}$$

Proof. The proof is immediate from Lemma 2.3 and Theorem 2.4. □

Remark. These results are novel in that they provide an exact expression (in two different forms!) for $\operatorname{sep}(M, -M^T) \equiv \min((\|M^T Z + ZM\|)/\|Z\|)$, where M is any stable matrix in $\mathbb{R}^{n\times n}$ [20]. This is because $\operatorname{sep}(M, -M^T) = \min(\|\Omega(Z)\|/\|Z\|) = \min(\|Y\|/\|\Omega^{-1}(Y)\|) = 1/\|\Omega^{-1}\|$, where $\Omega(Z) \equiv M^T Z + ZM$. From the preceding arguments, $\operatorname{sep}(M, -M^T) = 1/\|H\|$, where $M^T H + HM = -I$; and $\operatorname{sep}(M, -M^T) = 1/D^2$, where $D = \max_{\|x_0\|=1} \|x\|_{L_2}$ for solutions to $\dot{x} = Mx$, $x(0) = x_0$. This is important in another respect because it means that the damping measure, $D = D(M)$ is nicely stable: Stewart has shown [20, Thm. 4.6] that

$$\operatorname{sep}(M, -M^T) - 2\|E\| \leqq \operatorname{sep}(M + E, -(M+E)^T) \leqq \operatorname{sep}(M, -M) + 2\|E\|.$$

Thus, $D(M)/\sqrt{1 + 2\|E\| D^2(M)} \leqq D(M+E) \leqq D(M)/\sqrt{1 - 2\|E\| D^2(M)}$, where we assume that E is small enough that $M + E$ is stable and $2\|E\| D^2(M) < 1$.

From Theorem 2.4 we see that large norms for H_0, H_1, and H_2, which are indicative of poor damping in the closed-loop dynamical system, mean that K_B is large. Consequently, X will be sensitive to small perturbations in A, F, and G. Conversely, if the norms of H_0, H_1, and H_2 are not large then the problem is well-conditioned. A mixture of large and small norms among the H_k indicates selective sensitivity. For example, if the closed-loop dynamic system (1.7) is poorly damped with respect to the identity as a weighting matrix but not with respect to X or X^2, so that $\|H_0\|$ is large relative to $\|H_1\|$ and $\|H_2\|$, then we can expect that X is more sensitive to variations in the sensor matrix G than in the open loop matrix A or the controller matrix F. Examples showing selective sensitivity are given in § 4.

Although Theorem 2.4 gives exact values for $\|\Omega^{-1}\|$ and $\|\Pi\|$ in terms of $\|H_0\|$ and $\|H_2\|$, the upper and lower bounds on $\|\Theta\|$ in (2.15) are not very useful unless

$$\|H_1\| \cong \|H_0\|^{1/2}\|H_2\|^{1/2}. \tag{2.17}$$

The next lemma shows that this approximate equality must hold whenever X is moderately well-conditioned with respect to inversion.

LEMMA 2.6. *Assume that* $X>0$ *and let* $K(X)=\|X\|\,\|X^{-1}\|$. *Then*

$$\left[\frac{\|H_0\|\,\|H_2\|}{K(X)}\right]^{1/2} \leqq \|H_1\| \leqq [\|H_0\|\,\|H_2\|]^{1/2}.$$

Proof. By (2.15) we need only show $\|H_0\|\,\|H_2\| \leqq K(X)\|H_1\|^2$. Let v be any unit vector in $\mathbb{R}^n$. Then,

$$v^T H_1 v = \int_0^\infty v^T e^{A_C^T t} X e^{A_C t} v\, dt \geqq \lambda_{\min}(X) \int_0^\infty v^T e^{A_C^T t} e^{A_C t} v\, dt = \lambda_{\min}(X) v^T H_0 v.$$

This gives $\|H_1\| \geqq \lambda_{\min}(X)\|H_0\| = \|H_0\|/\|X^{-1}\|$. Similarly,

$$v^T H_1 v = \int_0^\infty v^T e^{A_C^T t} X X^{-1} X e^{A_C t} v\, dt \geqq \lambda_{\min}(X^{-1}) v^T H_2 v.$$

This gives $\|H_1\| \geqq \lambda_{\min}(X^{-1})\|H_2\| = \|H_2\|/\|X\|$. Thus $\|H_1\|^2 \geqq \|H_0\|\,\|H_2\|/K(X)$. □

For all but one of the problems that we tested the approximate equality (2.17) was true to within about 15 percent—even for some problems with $K(X)$ large. However, the following example shows that $\|H_0\|^{1/2}\|H_2\|^{1/2}$ can be very much larger than $\|H_1\|$.

Example 1. Let

$$A=\begin{bmatrix}0 & \lambda\\ 0 & 0\end{bmatrix}, \quad B=\begin{bmatrix}0\\ 1\end{bmatrix}, \quad C=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}. \tag{2.18}$$

Then

$$X=\begin{bmatrix}\sqrt{1+2\lambda}/\lambda & 1\\ 1 & \sqrt{1+2\lambda}\end{bmatrix},$$

$$\|H_0\| \cong \sqrt{\lambda/8}, \quad \|H_1\| \cong 1, \quad \|H_2\| \cong \sqrt{(9/8)\lambda}.$$

For example, if $\lambda = 10^8$ then $1 \cong \|H_1\| \leqq \|H_0\|^{1/2}\|H_2\|^{1/2} \cong 6124$.

Fortunately, there is a simple procedure that can be used to estimate $\|\Theta\|$ accurately even when $\|H_1\| \ll \|H_0\|^{1/2}\|H_2\|^{1/2}$. To motivate this method, we briefly consider the related problem of finding the Frobenius norms of Ω^{-1}, Θ, and Π.

Let Vec M denote the vector formed by stacking the columns of a matrix M and define the Kronecker product of two matrices M and N by (see [7])

$$M \otimes N \equiv [M_{ij}N]. \tag{2.19}$$

The Frobenius norm of M is equal to the 2-norm of Vec M:

$$\|M\|_F = \|\text{Vec } M\|. \tag{2.20}$$

Also

$$\text{Vec}\,(MZN) = (N^T \otimes M)\,\text{Vec } Z. \tag{2.21}$$

Apply the Vec operator to the equation $A_C^T H + HA_C = Q$, to obtain

$$(A_C^T \otimes I + I \otimes A_C^T) \operatorname{Vec} H = \operatorname{Vec} Q. \tag{2.22}$$

This may be rewritten as

$$\operatorname{Vec} H = (A_C^T \otimes I + I \otimes A_C^T)^{-1} \operatorname{Vec} Q. \tag{2.23}$$

Since $H = \Omega^{-1}(Q)$, we get

$$\begin{aligned} \|\Omega^{-1}\|_F &\equiv \max \frac{\|\Omega^{-1}(Q)\|_F}{\|Q\|_F} = \max \frac{\|H\|_F}{\|Q\|_F} = \max \frac{\|\operatorname{Vec} H\|}{\|\operatorname{Vec} Q\|} \\ &= \|(A_C^T \otimes I + I \otimes A_C^T)^{-1}\|. \end{aligned} \tag{2.24}$$

Note in (2.24) that $\|\Omega^{-1}\|_F$ denotes the operator norm on Ω^{-1} induced by the Frobenius matrix norm, and is not the same as the Frobenius norm of the associated Kronecker representation of Ω^{-1}. From (2.24) we see that we could evaluate $\|\Omega^{-1}\|_F$ by constructing $(A_C^T \otimes I + I \otimes A_C)^{-1}$ and taking its 2-norm (see [8] or [2]). However, because $A_C^T \otimes I + I \otimes A_C^T$ is of order n^2, this is usually not a practical procedure. Instead, we use the fact that if $M \equiv A_C^T \otimes I + I \otimes A_C^T$ then $\|M^{-1}\| = \sigma_{\max}^{1/2}((M^{-1})^T M^{-1})$.

Thus $\|M^{-1}\|$ can be estimated by the inverse power method [5]. That is, given $v = \operatorname{Vec}(Q) \neq 0$, solve for $\tilde{\omega}$ and ω in

$$M\tilde{\omega} = v, \qquad M^T \omega = \tilde{\omega}.$$

Unless v is poorly chosen, $\|M^{-1}\| \cong \sqrt{\|\omega\|/\|v\|}$. By recycling repeatedly with v reset to ω, this estimate converges to $\|M^{-1}\|$—unless the initial v is orthogonal to the eigenspace E_λ of $(M^T)^{-1}M^{-1}$ corresponding to $\lambda = \|M^{-1}\|^2$ (see [5]).

An equivalent version of this method avoids the use of Kronecker forms: given $Q \neq 0$, solve for $\tilde{H}$ and H in

$$A_C^T \tilde{H} + \tilde{H} A_C = Q, \tag{2.25}$$

$$A_C H + H A_C^T = \tilde{H}. \tag{2.26}$$

Then $\|\Omega^{-1}\|_F \cong \sqrt{\|H\|_F/\|Q\|_F}$ and recycling with Q reset to H improves the estimate [3].

The Frobenius norm of Θ can be found in a similar way. Suppose that $\Theta(Q) = H$, that is, $A_C^T H + HA_C = Q^T X + XQ$. Apply the Vec operator to get

$$(A_C^T \otimes I + I \otimes A_C^T) \operatorname{Vec} H = (I \otimes X) \operatorname{Vec} Q + (X \otimes I) \operatorname{Vec} Q^T.$$

The components of $\operatorname{Vec} Q^T$ are just a permutation of the components of $\operatorname{Vec} Q$:

$$\operatorname{Vec} Q^T = U \operatorname{Vec} Q \tag{2.27}$$

where U is a permutation matrix [7] with $U^T = U$. Then

$$\operatorname{Vec} H = (A_C^T \otimes I + I \otimes A_C^T)^{-1}(I \otimes X + (X \otimes I)U) \operatorname{Vec} Q \tag{2.28}$$

and

$$\|\Theta\|_F = \|(A_C \otimes I + I \otimes A_C^T)^{-1}(I \otimes X + (X \otimes I)U)\|. \tag{2.29}$$

The same type of argument gives

$$\|\Pi\|_F = \|(A_C \otimes I + I \otimes A_C^T)^{-1}(X \otimes X)\|. \tag{2.30}$$

To estimate $\|\Theta\|_F$, the inverse power method takes the following form. Given $\tilde{Q} \neq 0$, set $Q = \tilde{Q}^T X + X\tilde{Q}$, solve (2.25), (2.26), and let $W = 2XH$. Then $\|\Theta\|_F \cong \sqrt{\|W\|_F/\|\tilde{Q}\|_F}$, and we can recycle with $\tilde{Q}$ set equal to W. To estimate $\|\Pi\|_F$, set $Q = X\tilde{Q}X$, solve (2.25), (2.26), and let $Z = XHX$ to get $\|\Pi\|_F \cong \sqrt{\|Z\|_F/\|\tilde{Q}\|_F}$; recycle with $\tilde{Q} = Z$.

These results can be related to the 2-norms of Ω^{-1}, Θ, and Π by exploiting the well-known inequality [18]

$$\|M\| \leqq \|M\|_F \leqq \sqrt{n}\,\|M\|, \tag{2.31}$$

for any $n \times n$ matrix M. If L is either of the operators Ω^{-1}, Θ, or Π, then

$$\frac{1}{\sqrt{n}}\frac{\|L(Q)\|}{\|Q\|} \leqq \frac{\|L(Q)\|_F}{\|Q\|_F} \leqq \sqrt{n}\,\frac{\|L(Q)\|}{\|Q\|}, \tag{2.32}$$

and

$$\frac{\|L\|}{\sqrt{n}} \leqq \|L\|_F \leqq \sqrt{n}\,\|L\|. \tag{2.33}$$

This means that if Q maximizes the 2-norm ratio $\|L(Q)\|/\|Q\|$, then Q will nearly maximize the Frobenius norm ratio $\|L(Q)\|_F/\|Q\|_F$, and vice versa. For example, by Theorem 2.4, $\|\Omega^{-1}(Q)\|/\|Q\|$ is maximized by $Q = I$, so $\|\Omega^{-1}\|_F \cong \|\Omega^{-1}(I)\|_F/\|I\|_F = \|H_0\|_F/\sqrt{n}$. Similarly, $\|\Pi\|_F \cong \|\Pi(I)\|_F/\|I\|_F = \|H_2\|_F/\sqrt{n}$. If better estimates are required then the inverse power method can be used, starting with Q (or $\tilde{Q}$) equal to the identity.

This also suggests a means of estimating the 2-norm of Θ when $\|H_1\| \ll \|H_0\|^{1/2}\|H_2\|^{1/2}$: Use the inverse power method to find $\tilde{Q} \neq 0$ such that $\|\Theta(\tilde{Q})\|_F/\|\tilde{Q}\|_F$ is nearly maximized. Then $\|\Theta(\tilde{Q})\|/\|\tilde{Q}\|$ provides a good estimate for $\|\Theta\|$.

We found that for all the problems we tested, one cycle of the inverse power method, starting with $\tilde{Q} = I$, gave excellent estimates for $\|\Theta\|$. More specifically, set $Q = 2X$, solve (2.25), (2.26), and let $W = 2XH$. Define

$$H_1^{(1)} \equiv \Theta\left(\frac{W}{\|W\|}\right). \tag{2.34}$$

Then $\|H_1^{(1)}\| \cong \|\Theta\|$ and $\|H_1^{(1)}\|$ gives a lower bound on $\|\Theta\|$:

$$\|H_1^{(1)}\| = \frac{\|\Theta(W)\|}{\|W\|} \leqq \max \frac{\|\Theta(Q)\|}{\|Q\|} = \|\Theta\|.$$

If we apply this procedure to Example 1, then we find

$$\sqrt{\lambda/2} \cong \|H_1^{(1)}\| \leqq \|\Theta\| \leqq \sqrt{\|H_0\|\,\|H_2\|} \cong \sqrt{3\lambda/2}.$$

Thus $\|H_1^{(1)}\|$ is within $\sqrt{3}$ of $\|\Theta\|$ for this problem.

The preceding suggests that we define upper and lower condition estimates for the 2-norm as

$$K_U \equiv \frac{\|H_0\|\,\|G\| + 2[\|H_0\|\,\|H_2\|]^{1/2}\|A\| + \|H_2\|\,\|F\|}{\|X\|}, \tag{2.35}$$

$$K_L \equiv \frac{\|H_0\|\,\|G\| + \|H_1^{(1)}\|\,\|A\| + \|H_2\|\,\|F\|}{\|X\|}. \tag{2.36}$$

By Theorem 2.4,

$$K_L \leqq K_B \leqq K_U. \tag{2.37}$$

By Theorem 2.2,

$$\frac{K_L}{3} \leqq K \leqq K_U. \tag{2.38}$$

Numerical tests on a large class of problems (see § 4) show that K_L and K_U are usually very close. If desired, more refined results can be obtained by replacing the term $\|H_1^{(1)}\|$ in (2.36) with the norms of higher inverse power cycle iterates.

We conclude this section by giving a lower bound on $\|H_0\|$ in terms of the stability radius of the closed-loop matrix. For any matrix $M \in R^{m\times m}$, the stability radius of M is the norm of the smallest perturbation ΔM, which makes $M+\Delta M$ unstable (see [11], [14], [21]):

$$\rho_s(M) \equiv \min\{\|\Delta M\| \mid M+\Delta M \text{ has an eigenvalue } \lambda \text{ with } \mathrm{Re}(\lambda) \geqq 0\}. \tag{2.39}$$

LEMMA 2.7. *Let* $A_C^T H_0 + H_0 A_C = -I$. *Then*

$$\frac{1}{2\rho_s(A_C)} \leqq \|H_0\|. \tag{2.40}$$

Proof. See [9] for the proof. □

From this lemma we see that if the closed-loop matrix $A-FX$ can be made unstable by a small perturbation, then $\|H_0\|$ will be large and the problem of finding X is ill-conditioned. This complements a study by Kenney and Laub [14] in which it is shown that for companion systems

$$A = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ & & \ddots & \vdots \\ & & & 1 \\ a_1 & a_2 & \cdots & a_n \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad C = I \tag{2.41}$$

where a_1 does not grow faster than c^n for some constant $c>1$, we have that $\rho_s(A-BB^TX)$ decays to zero exponentially fast as n increases. We can thus expect that X is very sensitive to small changes in A, B, and C for companion form systems of high dimension.

3. Sensitivity of the differential Riccati equation. We now turn to the sensitivity of the solution to the differential Riccati equation, which although important has apparently not been treated in the literature.

For convenience, we transform the Riccati problem (1.4) with terminal condition into an initial value Riccati problem. Let $X(t) \equiv P(t_1 - t)$; then for $0 \leqq t \leqq t_1$

$$\dot{X}(t) = G + A^T X(t) + X(t)A - X(t)FX(t), \qquad X(0) = P_1 \equiv X_0. \tag{3.1}$$

Under this transformation, $P(t) = X(t_1 - t)$ and the closed-loop dynamical system (1.5) becomes

$$\dot{x}(t) = (A - FX(t_1 - t))x(t), \quad x(0) = x_0, \quad 0 \leqq t \leqq t_1. \tag{3.2}$$

As in the algebraic case, it is convenient to define the damping of the closed-loop dynamical system (3.2) with respect to the weighting matrices X^k for $k = 0, 1, 2$:

$$D_k(t_1) \equiv \max_{\|x_0\|=1} \|x\|_{L_2[0,t_1],X^k} \equiv \max_{\|x_0\|=1} \left[\int_0^{t_1} x^T(t) X^k(t_1 - t) x(t)\, dt\right]^{1/2} \tag{3.3}$$

where the maximum is taken over solutions, x to (3.2).

To avoid confusion, we will use a "+" subscript or superscript to denote quantities related to the algebraic Riccati equation (e.g., X_+ for the solution to (1.6), K_B^+ for (2.8), etc.).

Let $\tilde{X}_0 = \tilde{X}_0^T \geqq 0$ be a perturbation of X_0 and set $\Delta X_0 \equiv \tilde{X}_0 - X_0$. Let $\tilde{A}$, $\tilde{F}$, and $\tilde{G}$ be perturbations of A, F, and G with $\Delta A = \tilde{A} - A$, $\Delta F = \tilde{F} - F$, and $\Delta G = \tilde{G} - G$. As in § 2, we assume that $\tilde{F}$ and $\tilde{G}$ are symmetric and positive definite. Let $\tilde{X} = \tilde{X}(t)$ denote the solution to (3.1) for $\tilde{A}$, $\tilde{F}$, $\tilde{G}$, and $\tilde{X}_0$ and set

$$\Delta X(t) \equiv \tilde{X}(t) - X(t). \tag{3.4}$$

For $0 \leqq t \leqq t_1$, define the differential Riccati condition numbers

$$K_\delta(t) \equiv \sup \left\{ \frac{\|\Delta X(t)\|}{\delta \|X(t)\|} \right\}, \qquad K(t) \equiv \lim_{\delta \to 0} K_\delta(t). \tag{3.5}$$

Here the supremum is taken over the set $\|\Delta W\| \leqq \delta \|W\|$ with $W = A$, F, G, and X_0 such that $F + \Delta F$ and $G + \Delta G$ are symmetric and positive semidefinite. We assume throughout that $\|X(t)\| > 0$; obvious modifications in terms of absolute rather than relative condition numbers apply to the case where $\|X(t)\|$ is very small or zero.

Bounds on $K(t)$ can be found by expanding (3.1) for the perturbed matrices. For $X = X(t)$, let $A_C = A - FX$ and define

$$M_1 \equiv \Delta G + \Delta A^T X + X \Delta A - X \Delta F X, \tag{3.6}$$

$$M_2 \equiv (\Delta A - \Delta F X)^T \Delta X + \Delta X (\Delta A - \Delta F X) - \Delta X (F + \Delta F) \Delta X. \tag{3.7}$$

Using $\tilde{A}$, $\tilde{F}$, $\tilde{G}$, and $\tilde{X}$ in (3.1), we obtain

$$\Delta \dot{X} = A_C^T \Delta X + \Delta X A_C + M_1 + M_2, \quad \Delta X(0) = \Delta X_0, \quad 0 \leqq t \leqq t_1. \tag{3.8}$$

To find the analogue of Ω^{-1} in (2.4), let Φ satisfy

$$\dot{\Phi}(t) = \Phi(t) A_C(t), \quad \Phi(0) = I, \quad 0 \leqq t \leqq t_1. \tag{3.9}$$

Define

$$\Omega_t^{-1}(Z) \equiv -\int_0^t \Phi^T(t) \Phi^{-T}(s) Z(s) \Phi^{-1}(s) \Phi(t)\, ds \tag{3.10}$$

for any continuous matrix function $Z = Z(s)$, $s \in [0, t]$.

By variation of parameters [6], ΔX in (3.8) can be written as

$$\Delta X(t) = \Phi^T(t) \Delta X_0 \Phi(t) - \Omega_t^{-1}(\Delta G) - \Theta_t(\Delta A) + \Pi_t(\Delta F) + \Omega_t^{-1}(M_2) \tag{3.11}$$

where

$$\Theta_t(Z) \equiv \Omega_t^{-1}(Z^T X + XZ), \qquad \Pi_t(Z) \equiv \Omega_t^{-1}(XZX). \tag{3.12}$$

Since the first-order terms in (3.11) depend on the constant matrices ΔA, ΔF, ΔG, and ΔX_0, we define the restricted operator norm, $\|R\| \equiv \max (\|R(Z)\| / \|Z\|)$ for $R = \Omega_t^{-1}$, Θ_t, or Π_t with Z restricted to nonzero constant matrices in $\mathbb{R}^{n \times n}$.

From (3.11),

$$\begin{aligned} \|\Delta X(t)\| \leqq{}& \|\Phi(t)\|^2 \|\Delta X_0\| + \|\Omega_t^{-1}\| \|\Delta G\| + \|\Theta_t\| \|\Delta A\| + \|\Pi_t\| \|\Delta F\| \\ &+ \int_0^t \|\Phi^{-1}(s)\Phi(t)\|^2 \, ds [2(\|\Delta A\| + \|\Delta F\| \|X\|_t) \|\Delta X\|_t \\ &+ (\|F\| + \|\Delta F\|) \|\Delta X\|_t^2] \end{aligned}$$

where $\|\cdot\|_t$ denotes the max norm, $\|Z\|_t \equiv \max_{0\leq s\leq t}\|Z(s)\|$ for any continuous matrix function Z. This inequality suggests that we define the Byers-type approximate condition number

$$K_B(t) \equiv \frac{\|\Phi(t)\|^2\|X_0\| + \|\Omega_t^{-1}\|\,\|G\| + \|\Theta_t\|\,\|A\| + \|\Pi_t\|\,\|F\|}{\|X(t)\|}. \tag{3.13}$$

We then have the analogue of Theorem 2.2.

THEOREM 3.1. *For $0\leq t\leq t_1$, let $K(t)$, $K_B(t)$ be defined by* (3.5) *and* (3.13), *and assume that $X(t)$ satisfies* (3.1) *with $\|X(t)\|\neq 0$. Then*

$$\frac{K_B(t)}{4} \leq K(t) \leq K_B(t). \tag{3.14}$$

Proof. See Appendix 2 for the proof. □

This theorem shows that $K_B(t)$ gives a reasonable estimate of $K(t)$. The connection between conditioning and damping is established by using the analogues of H_k in (2.16): let $H_k = H_k(t)$ be the solution to

$$\dot{H}_k = A_C^T H_k + H_k A_C + X^k, \quad H_k(0) = 0, \quad k = 0, 1, 2. \tag{3.15}$$

LEMMA 3.2. *Let x, D_k, and H_k satisfy* (3.2), (3.3), *and* (3.15), *respectively. Then*

$$x_0^T H_k(t_1) x_0 = \|x\|^2_{L_2[0,t_1],X^k} \equiv \int_0^{t_1} x^T(t) X^k(t_1 - t) x(t)\, dt \tag{3.16}$$

and

$$\|H_k(t_1)\| = D_k^2(t_1) \quad \textit{for } k = 0, 1, 2. \tag{3.17}$$

Proof. Use the fact that $x(t) = \Phi^{-1}(t_1 - t)\Phi(t_1)x_0$ and

$$H_k(t) = \int_0^t \Phi^T(t)\Phi^{-T}(s) X^k(s) \Phi^{-1}(s)\Phi(t)\, ds, \qquad 0\leq t\leq t_1 \tag{3.18}$$

where Φ satisfies (3.9). Now proceed as in the proof of Lemma 2.3. □

From (3.18),

$$\Omega_t^{-1}(I) = -H_0(t), \quad \Theta_t(I) = -2H_1(t), \quad \Pi_t(I) = -H_2(t) \tag{3.19}$$

so that $\|H_0(t)\| \leq \|\Omega_t^{-1}\|$, $2\|H_1(t)\| \leq \|\Theta_t\|$, and $\|H_2(t)\| \leq \|\Pi_t\|$. This suggests the following analogues of Theorem 2.4.

THEOREM 3.3. *For H_k, Ω_t^{-1}, θ_t, and Π_t as in* (3.15), (3.10), *and* (3.12), *respectively,*

$$\|\Omega_t^{-1}\| = \|H_0(t)\|, \tag{3.20}$$

$$\|\Pi_t\| = \|H_2(t)\|, \tag{3.21}$$

$$2\|H_1(t)\| \leq \|\Theta_t\| \leq 2\|H_0(t)\|^{1/2}\|H_2(t)\|^{1/2}. \tag{3.22}$$

Proof. The proof is similar to the proof of Theorem 2.4. □

Because of the parallels between the algebraic and differential problems, the remarks following Theorem 2.4 also apply to Theorem 3.3 with only slight modifications, the main difference arising from the presence of the term $\|\Phi(t)\|^2\|X_0\|$ in (3.13). This term represents the main contribution to the error for t small because $H_k(0) = 0$, but as we will see, $\|\Phi(t)\| \to 0$ as $t\to\infty$ so that its influence is less important for large t. The norm of $\Phi(t)$ can be interpreted as a damping measure for the linear dynamical system (3.9) rather than (3.2).

For $X_0 = X_0^T \geqq 0$, $\lim_{t\to\infty} X(t) = X_+$, (see [4]) and $\lim_{t\to\infty} \dot{X}(t) = 0$. In this case, the differential relation (1.4) for X tends to the algebraic relation (1.6). Thus, it is natural to ask how the sensitivity estimates for the differential problem compare with those of the algebraic problem as $t \to \infty$. The next two lemmas show that the two sensitivity estimates are asymptotically equal, under the assumption that $\|H_0(t)\|$ is uniformly bounded for $0 \leqq t < \infty$.

LEMMA 3.4. *Let $X = X(t)$ satisfy* (3.1) *with $X_0 = X_0^T \geqq 0$. Let Φ satisfy* (3.9). *Then* $\lim_{t\to\infty} \Phi(t) = 0$.

Proof. See Appendix 2 for the proof. □

LEMMA 3.5. *Let Ω^{-1}, Θ, Π, and H_k be given by* (2.4), (2.7), *and* (2.11). *Let Ω_t^{-1}, θ_t, Π_t, and $H_k(t)$ be defined by* (3.10), (3.12), *and* (3.15). *If $\|H_0(t)\|$ is uniformly bounded over $0 \leqq t < \infty$, then*

$$\lim_{t\to\infty} H_k(t) = H_k. \tag{3.23}$$

Moreover, for any constant matrix Z,

$$\lim_{t\to\infty} \Omega_t^{-1}(Z) = \Omega^{-1}(Z), \tag{3.24}$$

$$\lim_{t\to\infty} \Theta_t(Z) = \Theta(Z), \tag{3.25}$$

and

$$\lim_{t\to\infty} \Pi_t(Z) = \Pi(Z). \tag{3.26}$$

Consequently, $\lim_{t\to\infty} K_B(t) = K_B^+$.

Proof. See Appendix 2 for the proof. □

Remark. Since $A_C(t) \to A_C^+$, which is stable, the assumption that $\|H_0(t)\|$ is uniformly bounded for $t \geqq 0$ seems reasonable; however, it is not clear how restrictive this assumption really is.

Just as in the algebraic problem, the possibility exists for $\|H_1(t)\| \ll \|H_0(t)\|^{1/2}\|H_2(t)\|^{1/2}$, in which case (3.22) does not provide much information about $\|\Theta_t\|$. In general this was not found to be a problem, but for Example 1 (2.18) of § 2, the gap increased with λ. For example, if $\lambda = 10^8$ and $X_0 = I$, then $\|H_1(10^{-4})\| = 0.724$, whereas

$$\|H_0(10^{-4})\|^{1/2}\|H_2(10^{-4})\|^{1/2} = 2432.0.$$

This problem can be handled as in § 2 by using the Kronecker forms associated with the Frobenius norm of Θ_t. For a given constant matrix Z,

$$\Theta_t(Z) = -\int_0^t \Phi^T(t)\Phi^{-T}(s)(Z^T X(s) + X(s)Z)\Phi^{-1}(s)\Phi(t)\, ds. \tag{3.27}$$

Using the Vec operator, we obtain

$$\operatorname{Vec}(\Theta_t(Z)) = L(t)\operatorname{Vec}(Z) \tag{3.28}$$

where

$$L(t) \equiv -\Phi^T(t) \otimes \Phi(t) \int_0^t (\Phi^{-T}(s) \otimes \Phi^{-T}(s))(I \otimes X(s) + (X(s) \otimes I)U)\, ds. \tag{3.29}$$

(The permutation matrix U is symmetric [7] and satisfies $U\,\mathrm{Vec}\,(Z^T)=\mathrm{Vec}\,(Z)$.) From (3.28),

$$\|\Theta_t\|_F=\max\frac{\|\Theta_t(Z)\|_F}{\|Z\|_F}=\max\frac{\|\mathrm{Vec}\,\Theta_t(Z)\|}{\|\mathrm{Vec}\,Z\|}=\max\frac{\|L(t)\,\mathrm{Vec}\,Z\|}{\|\mathrm{Vec}\,Z\|}=\|L(t)\|. \tag{3.30}$$

Thus the 2-norm of $L(t)$ is equal to the Frobenius norm (restricted to constant Z) of Θ_t. Moreover, from (3.29)

$$\dot{L}=(A_C^T\otimes I+I\otimes A_C^T)L+I\otimes X+(X\otimes I)U,\qquad L(0)=0. \tag{3.31}$$

We could integrate (3.31) to obtain $L(t)$ and then find $\|L(t)\|$ to get $\|\Theta_t\|_F$. In this case, an estimate of the 2-norm of Θ_t can be found by using the singular vector v of $L(t)$ corresponding to $\sigma_{\max}(L(t))$. Let $V=\mathrm{Unvec}\,(v)$, that is $\mathrm{Vec}\,(V)=v$, then $\|L(t)\|=\|L(t)v\|/\|v\|$ implies that $\|\Theta_t\|_F=\|\Theta_t(V)\|_F/\|V\|_F$. Thus by (2.32), we get the 2-norm estimate $\|\Theta_t\|\cong\|\Theta_t(V)\|/\|V\|$. In fact, by (2.33) we have $\|\Theta_t(V)\|/\|V\|\leqq\|\Theta_t\|\leqq\sqrt{n}(\|\Theta_t(V)\|/\|V\|)$.

This procedure is numerically expensive since $L(t)$ is $n^2\times n^2$. This suggests a power method approach like that of § 2, in which we avoid explicitly constructing $L(t)$. This method is only partially successful, because the nature of $L^T(t)$ makes it hard to avoid using Kronecker forms.

Given $\omega=\mathrm{Vec}\,(W)$, we can form $L(t)\omega$ by integrating the system

$$\dot{H}=A_C^TH+HA_C+W^TX+XW,\qquad H(0)=0 \tag{3.32}$$

and setting $L(t)\omega=\mathrm{Vec}\,(H(t))$. This follows from the (3.31), since $F\equiv L\omega$ satisfies

$$\dot{F}=(A_C^T\otimes I+I\otimes A_C^T)F+(I\otimes X+(X\otimes I)U)\omega.$$

However, $g\equiv L^T\omega$ satisfies $\dot{g}=L^T(A_C\otimes I+I\otimes A_C)\omega+(I\otimes X+U(X\otimes I))\omega$, which does not in general reduce to a simple $n\times n$ matrix differential equation for $\mathrm{Unvec}\,(L^T\omega)$, unless L^T and $(A_C\otimes I+I\otimes A_C)$ commute. In this case $\mathrm{Vec}\,(L^Tq)=H$ where

$$\dot{H}=A_CH+HA_C^T+X(W+W^T),\qquad H(0)=0.$$

This problem needs more research, but fortunately there is a procedure that seems to work well and is much less expensive than evaluating $L(t)$ directly. Let $\tilde{W}$ be given by one cycle of the inverse power method for estimating $\|\Theta\|_F$ as in (2.34). Let $H_1^{(1)}(t)$ satisfy (3.32) with $W=\tilde{W}/\|\tilde{W}\|$. Then $\|H_1^{(1)}(t)\|=\|\Theta_t(\tilde{W})\|/\|\tilde{W}\|\leqq\|\Theta_t\|$, and we have found that $\|H_1^{(1)}(t)\|\cong\|\Theta_t\|$. The rationale for this choice of $\tilde{W}$ is that $\|\Theta(\tilde{W})\|/\|\tilde{W}\|\cong\|\Theta\|$ as discussed in § 2, and that $\Theta_t\to\Theta$ as $t\to\infty$. Thus there is a good possibility that $\tilde{W}$ nearly maximizes the ratio $\|\Theta_t(\tilde{W})\|/\|\tilde{W}\|$, especially for large t.

We now define

$$K_U(t)\equiv\frac{\|\Phi(t)\|^2\|X_0\|+\|H_0(t)\|\,\|G\|+2\|H_0(t)\|^{1/2}\|H_2(t)\|^{1/2}\|A\|+\|H_2(t)\|\,\|F\|}{\|X(t)\|}, \tag{3.33}$$

$$K_L(t)\equiv\frac{\|\Phi(t)\|^2\|X_0\|+\|H_0(t)\|\,\|G\|+\|H_1^{(1)}(t)\|\,\|A\|+\|H_2(t)\|\,\|F\|}{\|X(t)\|}. \tag{3.34}$$

By (3.13) and Theorems 3.1 and 3.3, $K_L(t)\leqq K_B(t)\leqq K_U(t)$, $K_L(t)/4\leqq K(t)\leqq K_U(t)$. The asymptotic convergence results of Lemma 3.5 give $\lim_{t\to\infty}K_U(t)=K_U^+$ and $\lim_{t\to\infty}K_L(t)=K_L^+$.

4. Numerical results. By considering seven basic problems from [1], [9], and [15] with various parameter values, we tested 21 examples for the algebraic equation and seven for the differential equation. A representative subset of our results are presented below in such a way as to illustrate several points.

For example, the upper and lower bounds, K_U in (2.35) and K_L in (2.36) for the true condition number, K in (2.2), were farthest apart for the first example (see Table 1). The same applies to the time-dependent upper and lower bounds, $K_U(t)$ in (3.33) and $K_L(t)$ in (3.34) for $K(t)$ in (3.5), (see Table 2). That these upper and lower bounds are relatively close, even for this rather extreme example, is encouraging.

Another point we want to make with these results is that selective sensitivity can be easily detected. As discussed in § 2, the ratios $\|H_0\|\,\|G\|/\|X\|$, $\|H_1^{(1)}\|\,\|A\|/\|X\|$, and $\|H_2\|\,\|F\|/\|X\|$, (see (2.11) and (2.34)) measure, respectively, the sensitivity of X with respect to perturbations in the sensor matrix G, the state matrix A, and the actuator matrix F. Example 1 below illustrates state matrix sensitivity while Example 2 shows that actuator sensitivity dominates the condition number of companion form systems as the state dimension, n increases. The same example shows that the bound $1/2\rho_s(A_C) < \|H_0\|$, in Lemma 2.7, which relates the closed-loop stability radius to conditioning, can be very conservative. Example 3 shows that a problem can be sensitive in all three coefficient matrices. Lastly, Example 4 has the interesting feature that it is most sensitive with respect to sensor matrix perturbations for the algebraic Riccati problem, but for the differential problem most of the initial sensitivity ($t < 10$) is due to the actuator matrix. Except for Example 4, $F = BB^T$ and $G = C^TC$.

Example 1. Let

$$A = \begin{bmatrix} 0 & \lambda \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

TABLE 1
Condition measures for Example 1.

System parameter λ	Cond. no. (Frob. norm) K_B	Lower cond. no. (2-norm) K_L	Upper cond. no. (2-norm) K_U	Sensor matrix sensitivity $\frac{\|H_0\|\,\|G\|}{\|X\|}$	State matrix sensitivity $\frac{\|H_1^{(1)}\|\,\|A\|}{\|X\|}$	Actuator matrix sensitivity $\frac{\|H_2\|\,\|F\|}{\|X\|}$
10^0	3.7	3.8	4.2	0.6	1.6	1.6
10^1	7.4	7.3	11	0.3	6.3	0.8
10^2	53	52	89	0.3	51	0.8
10^4	5×10^3	5×10^3	9×10^3	0.3	5×10^3	0.8
10^8	5×10^7	5×10^7	9×10^7	0.3	5×10^7	0.8

TABLE 2
Convergence of time-dependent condition measures to steady state for $\lambda = 10^8$ *and* $X_0 = I$.

Time t	Lower cond. no. (2-norm) $K_L(t)$	Upper cond. no. (2-norm) $K_U(t)$	Initial matrix sensitivity $\frac{\|\Phi(t)\|^2\|X_0\|}{\|X(t)\|}$	Sensor matrix sensitivity $\frac{\|H_0(t)\|\,\|G\|}{\|X(t)\|}$	State matrix sensitivity $\frac{\|H_1^{(1)}(t)\|\,\|A\|}{\|X(t)\|}$	Actuator matrix sensitivity $\frac{\|H_2(t)\|\,\|F\|}{\|X(t)\|}$
0.0	1.0	1.0	1.0	0.0	0.0	0.0
10^{-7}	29	79	1.0	4×10^{-8}	28	4×10^{-6}
10^{-6}	3×10^3	7×10^3	1.0	3×10^{-7}	3×10^3	3×10^{-3}
10^{-5}	2×10^5	3×10^5	0.2	2×10^{-6}	2×10^5	0.8
10^{-4}	1×10^7	2×10^7	3×10^{-4}	7×10^{-3}	1×10^7	1.0
10^{-3}	5×10^7	9×10^7	4×10^{-10}	0.3	5×10^7	0.8

For this problem the upper and lower condition number bounds, K_U and K_L are relatively more separated than for any other problem we tested. (See also the discussion in § 2 concerning the gap between $\|H_1\|$ and $\|H_0\|^{1/2}\|H_2\|^{1/2}$.) Not surprisingly this problem exhibits sensitivity with respect to the state matrix A (compare column 6 with columns 5 and 7 in Table 1). This carries over to the differential problem. Table 2 shows the asymptotic convergence of the time-dependent condition measures to their steady state (algebraic) values.

Example 2. Let A, B, and C be given by (2.41), with $a_1 = a_2 = \cdots = a_n = 0$. This problem illustrates that high-order companion form systems are subject to actuator sensitivity (compare column 4 with 2 and 3 in Table 3). Columns 5 and 6 in Table 3 show that the bound $1/2\rho_s(A_C) < \|H_0\|$ in (2.40) can be very conservative. For related results, see [14].

Example 3 (Example 2 in [1]). Let A, B, and C be given by

$$A = \begin{bmatrix} -\varepsilon & 1 & 0 & 0 \\ -1 & -\varepsilon & 0 & 0 \\ 0 & 0 & \varepsilon & 1 \\ 0 & 0 & -1 & \varepsilon \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad C = B^T.$$

This problem was designed so that the closed-loop matrix A_C has eigenvalues that approach the imaginary axis as $\varepsilon \to 0$. This forces the stability radius $\rho_s(A_C)$ to go to zero and consequently by (2.40) the norm of H_0 becomes large. From Table 4, we see that X is sensitive to perturbations in A, B, or C.

Example 4 (From [12]). Let A, B, C, F, and G be given by

$$A = \begin{bmatrix} -0.0297 & 0.331 & -1.13 & 0.0 & 0.0 \\ -1.0 & -0.0042 & 0.128 & 0.0 & 1.0 \\ 0.0 & -0.0461 & -0.803 & 1.0 & 0.0 \\ 0.438 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 10^3 & 0 \\ 0 & 10^3 \end{bmatrix},$$

$$C = \begin{bmatrix} -0.00297 & 0.0331 & -0.0113 & 0.0 & 0.0 \\ 0.0 & 0.012048 & 0.021187 & 0.0 & 0.0 \\ 0.0 & 0.001265 & 0.05028 & 0.0 & 0.0 \end{bmatrix}, \quad F = BB^T,$$

$$G = C^T \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \varepsilon \end{bmatrix} C.$$

TABLE 3
Sensitivity measures for companion form systems.

System size n	Sensor matrix sensitivity $\frac{\|H_0\|\|G\|}{\|X\|}$	State matrix sensitivity $\frac{\|H_1^{(1)}\|\|A\|}{\|X\|}$	Actuator matrix sensitivity $\frac{\|H_2\|\|F\|}{\|X\|}$	Inverse stability radius $\frac{1}{2\rho_s(A_C)}$	Inverse Lyapunov norm $\|H_0\|$
5	0.65	4.5	39.0	15.6	40.0
10	0.65	9.2	8.2×10^3	65.0	9.5×10^3
15	0.65	14.0	2.2×10^5	1.4×10^3	2.6×10^6
20	0.64	19.0	6.3×10^8	2.4×10^3	7.8×10^8

TABLE 4
Sensitivity measures for Example 3.

Eigenvalue parameter ε	Sensor matrix sensitivity $\frac{\|H_0\|\|G\|}{\|X\|}$	State matrix sensitivity $\frac{\|H_1^{(1)}\|\|A\|}{\|X\|}$	Actuator matrix sensitivity $\frac{\|H_2\|\|F\|}{\|X\|}$
10^0	4.1	4.0	25
10^{-3}	4×10^6	2×10^6	4×10^6
10^{-5}	4×10^{10}	2×10^{10}	4×10^6
10^{-7}	4×10^{14}	2×10^{14}	4×10^6

This problem shows increasing sensitivity with respect to the sensor matrix G as ε decreases (see column 1 of Table 5). A very interesting feature of this problem is that for $X_0 = I$ and t near zero, the sensitivity of the differential Riccati equation (Table 6) is mostly due to the actuator matrix, F because of its large norm; $\|F\| = 10^6$. (See row 6, column 4 of Table 6.) Thus the sensitivity of the algebraic problem is not always a reliable guide to the sensitivity of the differential problem.

5. Conclusion. In this paper we have shown that the ideas developed by Byers in [2] for the algebraic Riccati equation can be sharpened and extended to norms other

TABLE 5
Algebraic sensitivity measures for Example 4.

Sensor parameter ε	Sensor matrix sensitivity $\frac{\|H_0\|\|G\|}{\|X\|}$	State matrix sensitivity $\frac{\|H_1^{(1)}\|\|A\|}{\|X\|}$	Actuator matrix sensitivity $\frac{\|H_2\|\|F\|}{\|X\|}$
10^0	10^3	9.9	10^2
10^{-1}	9×10^3	13	82
10^{-2}	2×10^4	13	82
10^{-3}	2×10^4	13	82

TABLE 6
Differential sensitivity for Example 4 *with* $\varepsilon = 10^{-3}$ *and* $X_0 = I$.

Time t	Sensor matrix sensitivity $\frac{\|H_0(t)\|\|G\|}{\|X(t)\|}$	State matrix sensitivity $\frac{\|H_1^{(1)}(t)\|\|A\|}{\|X(t)\|}$	Actuator matrix sensitivity $\frac{\|H_2(t)\|\|F\|}{\|X(t)\|}$	Initial matrix sensitivity $\frac{\|\Phi(t)\|^2\|X_0\|}{\|X(t)\|}$
0.0	0.0	0.0	0.0	1.0
10^{-5}	10^{-7}	10^{-4}	10^1	1.0
10^{-4}	10^{-6}	10^{-3}	10^2	1.0
10^{-3}	10^{-5}	10^{-2}	10^3	1.0
10^{-2}	10^{-2}	10^{-1}	10^4	1.0
10^{-1}	0.2	1.3	10^5	0.9
10^0	16	2.4	10^4	10^{-2}
10^1	59	13	6×10^3	7×10^{-4}
10^2	77	13	5×10^3	4×10^{-4}
10^3	6×10^3	13	83	6×10^{-8}

than the Frobenius norm. This extension is crucial from an interpretive point of view because use of the spectral norm allows an identification of the condition number of the algebraic Riccati equation with the damping properties of the closed-loop dynamical system. The resulting condition numbers are easily computed and provide a simple means of detecting selective sensitivity with respect to the sensor matrix, the open-loop state matrix or the actuator matrix. Moreover, this approach has the pleasant feature that it carries over to a completely parallel theory for the sensitivity of the differential Riccati equation, an area which to our knowledge has not been considered previously.

Appendix 1. Proofs of Theorems 2.2 and 2.4.

Proof of Theorem 2.2. Let $\|\Delta A\| \leqq \delta \|A\|$, $\|\Delta F\| \leqq \delta \|F\|$, and $\|\Delta G\| \leqq \delta \|G\|$ with $F+\Delta F$, $G+\Delta G$ symmetric and positive semidefinite. Then by (2.1), $\|\Delta X\| \leqq \delta \|X\| K_\delta$. Taking norms in (2.5),

$$\|\Delta X\| \leqq \delta(\|\Omega^{-1}\| \|G\| + \|\Theta\| \|A\| + \|\Pi\| \|F\|) + \delta^2 \|\Omega^{-1}\| K_\delta (2(\|A\| + \|F\| \|X\|)\|X\| + (1+\delta)\|F\| \|X\|^2 K_\delta).$$

Divide by $\|X\|\delta$ to get

$$\frac{\|\Delta X\|}{\delta \|X\|} \leqq K_B + \delta \|\Omega^{-1}\| K_\delta (2(\|A\| + \|F\| \|X\|) + (1+\delta)\|F\| \|X\| K_\delta).$$

From (2.1) we then obtain

$$K_\delta \leqq K_B + \delta \|\Omega^{-1}\| K_\delta (2(\|A\| + \|F\| \|X\|) + (1+\delta)\|F\| \|X\| K_\delta).$$

From Theorem 2.1, $K_\delta(A, F, G, \|\cdot\|_F)$ is finite for δ sufficiently small. Since all norms on $\mathbb{R}^{m\times m}$ are equivalent [18], we must have that $K_\delta(A, F, G, \|\cdot\|)$ is finite for δ small for the matrix 2-norm. But $K_\delta(A, F, G, \|\cdot\|)$ is nonincreasing as $\delta \to 0^+$ so

$$K(A, F, G, \|\cdot\|) = \lim_{\delta \to 0} K_\delta(A, F, G, \|\cdot\|) \leqq K_B(A, F, G, \|\cdot\|).$$

We now show that $\|\Omega^{-1}\| \|G\| \leqq K\|X\|$, $\|\Theta\| \|A\| \leqq K\|X\|$, and $\|\Pi\| \|F\| \leqq K\|X\|$; adding these inequalities gives $K_B \leqq 3K$, which completes the proof.

Let $\Delta A = 0$, $\Delta F = 0$, and $\Delta G = \delta \|G\| I$. Then $G+\Delta G$ is symmetric and positive definite as required in (2.1). Using these perturbation matrices in (2.5) gives $\Delta X = -\delta \|G\| \Omega^{-1}(I)$, so $\|\Delta X\| = \delta \|G\| \|\Omega^{-1}(I)\| = \delta \|G\| \|\Omega^{-1}\|$ by (2.12) and (2.13). From (2.1),

$$\|G\| \|\Omega^{-1}\| = \|\Delta X\|/\delta \leqq K_\delta \|X\|.$$

Letting $\delta \to 0$ gives $\|\Omega^{-1}\| \|G\| \leqq K\|X\|$ as desired. Similarly, $\Delta A = 0$, $\Delta F = \delta \|F\| I$, and $\Delta G = 0$ gives $\|\Theta\| \|F\| \leqq K\|X\|$, by (2.12) and (2.14). (Note that $\tilde{F} = F + \Delta F$ is symmetric and positive semidefinite in keeping with (2.1).)

For the operator norm defined by (2.9), standard compactness arguments show that there exists a matrix A_0, satisfying $\|A_0\| = 1$, $\|\Theta(A_0)\| = \|\Theta\|$. Now let $\Delta A = \delta \|A\| A_0$, $\Delta F = 0$, and $\Delta G = 0$ and after an argument similar to the one above we get $\|\Theta\| \|A\| \leqq K\|X\|$. □

Proof of Theorem 2.4. From (2.12), $\|H_0\| \leqq \|\Omega^{-1}\|$, $\|H_2\| \leqq \|\Pi\|$, and $2\|H_1\| \leqq \|\Theta\|$. Now, for $Z \in \mathbb{R}^{m\times m}$, let u and v be unit left and right singular vectors of $\Omega^{-1}(Z)$ such that

$$\|\Omega^{-1}(Z)\| = -u^T \Omega^{-1}(Z) v.$$

Then by (2.4), $\|\Omega^{-1}(Z)\| = \int_0^\infty u^T e^{A_C^T t} Z e^{A_C t} v \, dt \leqq \|Z\| \int_0^\infty \|e^{A_C t} u\| \, \|e^{A_C t} v\| \, dt$. Applying the Cauchy-Schwarz inequality, we obtain $\|\Omega^{-1}(Z)\| \leqq \|Z\| [\int_0^\infty \|e^{A_C t} u\|^2 \, dt]^{1/2} [\int_0^\infty \|e^{A_C t} v\|^2 \, dt]^{1/2}$. But $\int_0^\infty \|e^{A_C t} u\|^2 \, dt = u^T \int_0^\infty e^{A_C^T t} e^{A_C t} \, dt \, u = u^T H_0 u \leqq \|H_0\|$ because u is a unit vector. The same applies to v, so $\|\Omega^{-1}(Z)\| \leqq \|Z\| \, \|H_0\|^{1/2} \|H_0\|^{1/2} = \|Z\| \, \|H_0\|$. Thus $\|\Omega^{-1}\| \leqq \|H_0\|$. Similar arguments establish $\|\Pi\| \leqq \|H_2\|$ and $\|\Theta\| \leqq \|H_0\|^{1/2} \|H_2\|^{1/2}$. □

Appendix 2. Proofs of Theorem 3.1, Lemma 3.4, and Lemma 3.5.

Proof of Theorem 3.1. Let $t \in (0, \infty)$ be given. Because the right-hand side of (3.1) is a smooth function of its arguments, we may appeal to the theory of perturbations of initial conditions and parameters of ordinary differential equations (see [6, Thm. 7.5, Chap. 1]), to conclude that

$$\Delta X = \Delta X(A, F, G, X_0, \Delta A, \Delta F, \Delta G, \Delta X_0, t)$$

is a continuously differentiable function of its arguments. Moreover, for δ sufficiently small, with $\|\Delta A\|/\|A\|, \|\Delta F\|/\|F\|, \|\Delta G\|/\|G\|, \|\Delta X_0\|/\|X_0\| \leqq \delta$, there exists a constant $c < \infty$ such that $\|\Delta X\|_t \leqq \delta c$.

Using this in (3.11), we obtain

$$\|\Delta X(t)\| \leqq \|\Phi(t)\|^2 \delta \|X_0\| + \|\Omega_t^{-1}\| \delta \|G\| + \|\Theta_t\| \delta \|A\| + \|\Pi_t\| \delta \|F\| + \int_0^t \|\Phi^{-1}(s)\Phi(t)\|^2 \, dt [2(\|A\| + \|F\| \, \|X\|_t)c + (1+\delta)\|F\| c^2] \delta^2.$$

Thus $\|\Delta X(t)\|/\delta \|X(t)\| \leqq K_B(t) + N(t)\delta$, where

$$N(t) = \frac{\int_0^t \|\Phi^{-1}(s)\Phi(t)\|^2 \, dt}{\|X(t)\|} [2(\|A\| + \|F\| \, \|X\|_t)c + (1+\delta)\|F\| c^2].$$

So $K_\delta(t) \leqq K_B(t) + N(t)\delta$. Now $N(t)$ is finite by (1.10) and the assumption that $X(t) \neq 0$. Thus letting $\delta \to 0$ gives $K(t) = \lim_{\delta \to 0} K_\delta(t) \leqq K_B(t)$, which proves the right-hand side of (3.14).

Using arguments similar to those in the proof of Theorem 2.2, we can show that

$$\frac{\|\Omega_t^{-1}\| \, \|G\|}{\|X(t)\|} \leqq K(t), \qquad \frac{\|\Theta_t\| \, \|A\|}{\|X(t)\|} \leqq K(t),$$

$$\frac{\|\Pi_t\| \, \|F\|}{\|X(t)\|} \leqq K(t), \qquad \frac{\|\Phi(t)\|^2 \|X_0\|}{\|X(t)\|} \leqq K(t).$$

Adding these inequalities gives $K_B(t) \leqq 4K(t)$ and completes the proof. □

Proof of Lemma 3.4. This lemma has been proved in [13] for the special case $X_0 = 0$. For the general case write (3.9) as $\dot{\Phi}(t) = \Phi(t)(A - FX_+) + \Phi(t)F(X_+ - X(t))$. Since $A_C^+ \equiv A - FX_+ \in \mathbb{R}^{n \times n}$ is stable, the result follows from Theorem 8.1 on p. 92 and Problem 35 on p. 106 of [6], if we can show that for some $t_0 \geqq 0$,

$$\text{(A1)} \qquad \int_{t_0}^\infty t^n \|X_+ - X(t)\| \, dt < \infty.$$

To prove (A1), let $\Phi_+(t) = e^{A_C^+ t}$. Then we can use the representation (see [13])

$$\text{(A2)} \qquad X(t) - X_+ = \Phi_+^T(t - t_0) \left[I + (X(t_0) - X_+) \int_{t_0}^t \Phi_+(s - t_0) F \Phi_+^T(s - t_0) \, ds \right]^{-1} \cdot (X(t_0) - X_+) \Phi_+(t - t_0)$$

for any t_0 such that the inverse in (A2) exists.

Let $\Lambda(A_C^+)$ denote the set of eigenvalues of A_C^+ and define $\alpha(A_C^+) \equiv \max_{\lambda \in \Lambda(A_C^+)} R_e\lambda$. Since A_C^+ is stable, $\alpha = \alpha(A_C^+) < 0$, and there exists a constant c_0 such that $\|\Phi_+(t)\| \leqq c_0 e^{t\alpha/2}$. Hence, $\int_0^\infty \|\Phi_+(t)\|^2\, dt \leqq c_0^2 \int_0^\infty e^{\alpha t}\, dt = c_0^2/|\alpha| < \infty$. Since $\lim_{t\to\infty} X(t) = X_+$, we may find a $t_0 \geqq 0$ such that $\|X(t_0) - X_+\| \|F\| \int_{t_0}^\infty \|\Phi(t)\|^2\, dt < \frac{1}{2}$. This means that for all $t \geqq t_0$, the inverse in (A2) exists and

$$\text{(A3)} \qquad \left\| \left(I + (X(t_0) - X_+) \int_{t_0}^t \Phi_+(s - t_0) F \Phi^T(s - t_0)\, ds \right)^{-1} \right\| < 2.$$

Using (A3) in (A2) gives $\|X(t) - X_+\| \leqq 2c_0^2 \|X(t_0) - X_+\|\, e^{\alpha(t-t_0)}$, which establishes (A1) and completes the proof. □

Remark. The preceding actually shows that the decay rate for $\|\Phi(t)\|$ is asymptotically the same as the decay rate for $\|\Phi_+(t)\|$ (see [10, Thm. 5.4.1]).

Proof of Lemma 2.5. We show that $\lim_{t\to\infty} \Theta_t(Z) = \Theta(Z)$. The other asymptotic convergence proofs are very similar. From the definition of $\Theta_t(Z)$,

$$\dot{\Theta}_t(Z) = A_C^T \Theta_t(Z) + \Theta_t(Z) A_C + Z^T X + XZ, \qquad \Theta_0(Z) = 0.$$

This matrix differential equation can be written as a vector differential system: $\dot{Y}(t) = \mathscr{A}(t) Y(t) + S(t)$, where $Y_t(Z) = \text{Vec}\,(\Theta_t(Z))$, $\mathscr{A}(t) = I \otimes A_C^T + A_C^T \otimes I$, and $S(t) = \text{Vec}\,(Z^T X + XZ)$. Let $\mathscr{A}_+ = I \otimes (A_C^+)^T + (A_C^+)^T \otimes I$, and $S_+ = \text{Vec}\,(Z^T X_+ + X_+ Z)$. By Lemma 2.4, $\lim_{t\to\infty} \mathscr{A}(t) = \mathscr{A}_+$ and $\lim_{t\to\infty} S(t) = S_+$. Moreover, if we set $\alpha = \alpha(A_C^+)$ then there exists constants c_1 and c_2 such that by

$$\|\mathscr{A}(t) - \mathscr{A}_+\| = \|I \otimes (X_+ - X(t))F + F(X_+ - X(t)) \otimes I\| \leqq c_1 e^{t\alpha/2},$$

$$\|S(t) - S_+\| \leqq c_2 e^{t\alpha/2}.$$

Now $\mathscr{A}_+$ has eigenvalues of the form $\lambda = \mu + \nu$ where μ and ν are eigenvalues of A_C^+ (see [17]). Thus $\alpha(\mathscr{A}_+) = 2\alpha(A_C^+) = 2\alpha$. Hence there exists a constant c_3 such that $\|e^{\mathscr{A}_+ t}\| \leqq c_3 e^{\alpha t}$. Let $Y_+ \equiv \text{Vec}\,(\Theta(Z))$ and note that $\mathscr{A}_+ Y_+ = S_+$. Define $\Delta Y(t) \equiv Y(t) - Y_+$. Then $\Delta \dot{Y}(t) = \mathscr{A}_+ \Delta Y(t) + R(t)$, where $R(t) = (\mathscr{A}(t) - \mathscr{A}_+) Y(t) + S(t) - S_+$. The boundness of $\|H_0(t)\|$, together with $\|H_k(t)\| \leqq \|H_0(t)\|\, \|X^k\|_t$ for $k = 1, 2$ and $\|\Theta_t(Z)\| \leqq 2\|H_0(t)\|^{1/2} \|H_2(t)\|^{1/2}, \|Z\|$ ensures that $\|R(t)\| \leqq c_4 e^{t\alpha/2}$ for some constant c_4.

If we write $\Delta Y(t) = e^{\mathscr{A}_0 t} \Delta Y(0) + \int_0^t e^{\mathscr{A}_0(t-s)} R(s)\, ds$, we have

$$\|\Delta Y(t)\| \leqq c_3 e^{\alpha t} \|\Delta Y(0)\| + \int_0^t c_3 c_4 e^{\alpha(t-s)} e^{t\alpha/2}\, ds$$

$$\leqq e^{\alpha t} c_3 \|\Delta Y(0)\| + \frac{c_3 c_4 e^{t\alpha/2}}{|\alpha|} [1 - e^{\alpha t}].$$

Thus $\|\Delta Y(t)\| \to 0$ as $t \to 0$ and we must have $\Theta_t(Z) \to \Theta(Z)$. □

REFERENCES

[1] W. ARNOLD, *Numerical solution of algebraic matrix Riccati equations*, Tech. Paper TP6521, Naval Weapons Center, China Lake, CA, 1984.

[2] R. BYERS, *Numerical condition of the algebraic Riccati equation*, in Proc. Summer Research Conference, AMS Vol. 47, Contemporary Math., American Mathematical Society, Providence, RI, 1984, pp. 35-49.

[3] ———, *A* LINPACK*-style condition estimator for the equation* $AX - XB^T = C$, IEEE Trans. Automat. Control, 29 (1984), pp. 926-928.

[4] F. CALLIER AND J. WILLEMS, *Criterion for the convergence of the solution of the Riccati differential equation*, IEEE Trans. Automat. Control, 26 (1981), pp. 1232-1242.

[5] A. CLINE, C. B. MOLER, G. W. STEWART, AND J. H. WILKINSON, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368-375.

[6] E. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[7] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, Ellis Harwood Limited, John Wiley, New York, 1981.

[8] G. GOLUB, S. NASH, AND C. VAN LOAN, *A Hessenberg-Schur method for the problem, $AX + XB = C$*, IEEE Trans. Automat. Control, 24 (1979), pp. 909-913.

[9] G. HEWER AND C. KENNEY, *The sensitivity of the stable Lyapunov equation*, SIAM J. Control Optim., 26 (1988), pp. 321-344.

[10] E. HILLE, *Lectures on Ordinary Differential Equations*, Addison-Wesley, Reading, MA, 1969.

[11] D. HINRICHSEN AND A. PRITCHARD, *Stability radii of linear systems*, Systems Control Lett., 7 (1986), pp. 1-10.

[12] T. KAILATH, *Some new algorithms for recursive estimation in constant linear systems*, IEEE Trans. Inform. Theory, 19 (1973), pp. 750-760.

[13] T. KAILATH AND L. LJUNG, *The asymptotic behavior of constant-coefficient Riccati differential equations*, IEEE Trans. Automat. Control, 21 (1976), pp. 385-388.

[14] C. KENNEY AND A. LAUB, *Controllability and stability radii for companion form systems*, Math. of Control, Signals, Systems, 1 (1988), pp. 239-256.

[15] C. KENNEY AND R. LEIPNIK, *Numerical integration of the differential matrix Riccati equation*, IEEE Trans. Automat. Control, 30 (1985), pp. 962-970.

[16] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley Interscience, New York, 1972.

[17] P. LANCASTER, *Explicit solutions of linear matrix equations*, SIAM Rev., 12 (1970), pp. 544-566.

[18] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, 2nd ed., Academic Press, New York, 1985.

[19] J. RICE, *A Theory of Condition*, SIAM Numer. Anal., 3 (1966), pp. 287-310.

[20] G. STEWART, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727-764.

[21] C. VAN LOAN, *How near is a stable matrix to an unstable matrix?*, in Proc. Summer Research Conference, AMS Vol. 47 Contemporary Mathematics, American Mathematical Society, Providence, RI, 1984, pp. 465-478.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 70-96, January 1990

004

PIECEWISE LINEAR APPROXIMATION FOR HEREDITARY CONTROL PROBLEMS*

GEORG PROPST†

Abstract. This paper presents finite-dimensional approximations for linear retarded functional differential equations by use of discontinuous piecewise linear functions. The approximation scheme is applied to optimal control problems, when a quadratic cost integral must be minimized subject to the controlled retarded system. It is shown that the approximate optimal feedback operators converge to the true ones both in the case where the cost integral ranges over a finite time interval, as well as in the case where it ranges over an infinite time interval. The arguments in the last case rely on the fact that the piecewise linear approximations to stable systems are stable in a uniform sense. This feature is established using a vector-component stability criterion in the state space $\mathbb{R}^n \times L^2$ and the favorable eigenvalue behavior of the piecewise linear approximations.

Key words. hereditary control problem, piecewise linear approximation, uniform stability

AMS(MOS) subject classifications. 34K35, 65L60, 93C15

1. Introduction. Given $\phi^0 \in \mathbb{R}^n$ and $\phi^1 : [-h, 0] \to \mathbb{R}^n$, consider the retarded functional differential equation with constant coefficients

$$\dot{x}(t) = \sum_{k=0}^{p} A_k x(t-h_k) + \int_{-h}^{0} A_{01}(s) x(t+s)\, ds, \qquad t \geqq 0, \tag{1.1}$$
$$x(0) = \phi^0, \quad x = \phi^1 \quad \text{in } L^2(-h, 0; \mathbb{R}^n),$$

where $0 = h_0 < h_1 < \cdots < h_p = h$, $A_k \in \mathbb{R}^{n \times n}$, $k = 0, \cdots, p$, and $A_{01} \in L^2(-h, 0; \mathbb{R}^{n \times n})$. An equivalent abstract Cauchy problem $\dot{z}(t) = Az(t)$, $t \geqq 0$, $z(0) = (\phi^0, \phi^1)$ in the space $M^2 = \mathbb{R}^n \times L^2(-h, 0; \mathbb{R}^n)$ generates a strongly continuous semigroup. Approximations are constructed by restricting the problem to finite-dimensional subspaces $Z^N = \mathbb{R}^n \times Y^N \subseteq M^2$, defining appropriate generators A^N on Z^N.

Banks and Burns [1] used subspaces Y^N consisting of functions that are piecewise constant on the delay interval $[-h, 0]$. This is the well-known averaging approximation scheme. As an extension, Burns and Cliff [5] enlarged the subspaces to piecewise linear functions. In both papers, the approximating generators were constructed by forward difference methods. In [3] the approximations were obtained by projections onto subspaces of continuous splines being contained in the domain of A. Then Kappel and Salamon [13] introduced δ-type operators to define generators for a spline scheme whose adjoint semigroups converge strongly. These δ-type operators are specially constructed to approximate the differential operator A at the discrete delays, where the splines may be discontinuous, as are the functions in the domain of A^*.

In this paper, a new scheme is presented employing again subspaces of orthogonal piecewise linear functions as in [5], but using δ-type operators for the construction of the approximating generators. These operators are needed at each meshpoint, where the subspace functions may be discontinuous. In fact, the number of discontinuities

* Received by the editors March 9, 1987; accepted for publication (in revised form) March 24, 1989.
† Institut für Mathematik, Universität Graz, Elisabethstrasse 16, A-8010 Graz, Austria. This research was supported in part by National Aeronautics and Space Administration contracts NAS1-17070 and NAS1-18107 while the author was in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, Virginia 23665, and in part by the Fonds zur Förderung der wissenschaftlichen Forschung (Austria) under grant S3206 while the author was in residence at the Universität Graz, Graz, Austria in connection with his Ph.D. thesis.

increases with the order of the approximation, in contrast to the spline case. The resulting generators are completely different from those given in [5].

An application of the approximation schemes is the optimal control problem, when an integral ranging over $[0, T]$, quadratic in the trajectory and in the control, is to be minimized subject to the controlled delay equation. It was shown by Gibson [9] that, if $T<\infty$, the strong convergence of the semigroups and their adjoints yields convergence in norm of the optimal feedback operators. In this present work, strong convergence of the semigroups and their adjoints is proved using the Trotter–Kato Theorem as in [1] and [9]. Thereby, it is not necessary to assume absolute continuity of A_{01}, as did the proofs in [11]–[13].

In the case of the so-called infinite time horizon $T=\infty$, Gibson's approach relies on the assumption that a stable system is approximated by systems that are stable in a uniform sense. For the averaging scheme, this stability preservation property was established in [19], and in [11] for the Legendre-tau methods. In contrast, the spline schemes do not have this quality (see [14]). This is due to extraneous eigenvalues close to the imaginary axis. It is shown below that the eigenvalues of the present scheme converge to those of the delay equation and that exponential stability of our approximations is dominated by their $\mathbb{R}^n$-components. Thus, uniform preservation of exponential stability is proved with decay rates arbitrarily close to the decay rate of the hereditary system.

The matrices corresponding to the piecewise linear functions are banded and sparse, in contrast to the Legendre and spline methods. While the Legendre schemes [10], [12] exhibit high accuracy even for low-order approximation, the numerical efficiency of the present scheme is about the same as that of the first-order splines in [13], and superior to the averaging methods [1] as well as to those in [5].

Preliminarily, § 2 collects some facts on the semigroup generated by the uncontrolled system and on the linear quadratic hereditary control problem. Section 3 presents an approximation framework suited to § 4, where the piecewise linear scheme is developed. For the sake of transparency and brevity the presentation here is restricted to the one-delay case. The modifications necessary for the treatment of more general cases are briefly described in § 4.6. In § 4.1, the δ-type operators are defined and the projections onto the subspaces of piecewise linear functions are investigated. In § 4.2, the approximating generators and their adjoints are constructed and convergence results for the finite-time horizon problem are proved. In § 4.3, the matrix representations are given and an $\mathbb{R}^n$-component stability criterion is established. Section 4.4 investigates the eigenvalue behavior of the approximate systems when the order of approximation increases. In § 4.5, the uniform stability preservation is proved. Finally, in § 4.7 there is a brief discussion of the numerical tools needed for the implementation of the scheme on a computer and the results of three examples are tabulated.

2. The linear quadratic optimal control problem for hereditary systems. We give a brief outline of known results on the LQR problem. Proofs and a rigorous analysis may be found in [6], [8], and [9]. Consider the linear retarded functional differential equation

$$\dot{x}(t)=A_0x(t)+A_1x(t-r)+B_0u(t), \qquad t\geqq 0 \tag{2.1.1}$$

in $\mathbb{R}^n$, where $r>0$ is a fixed finite delay, A_0, A_1 are real $n\times n$ matrices, B_0 is a real $n\times m$ matrix, and $u(t)\in\mathbb{R}^m$.

Given $\phi=(\phi^0,\phi^1)\in M^2=\mathbb{R}^n\times L^2(-r,0;\mathbb{R}^n)$ and $u\in L^2_{\mathrm{loc}}(0,\infty;\mathbb{R}^m)$, there exists a unique solution $x(t;\phi,u)$ that is absolutely continuous with L^2-derivative on every

interval $[0, T]$ and that satisfies (2.1.1) for almost all $t \geqq 0$, and the initial condition

$$x(0; \phi, u) = \phi^0, \quad x(\cdot; \phi, u) = \phi^1 \quad \text{in } L^2(-r, 0; \mathbb{R}^n). \tag{2.1.2}$$

Defining $x_t: [-r, 0] \to \mathbb{R}^n$ by $x_t(s) = x(t+s)$ and the state at time t by

$$z(t; \phi, u) = (x(t; \phi, u), x_t(\phi, u)) \in M^2, \tag{2.2}$$

system (2.1) is converted to an abstract Cauchy problem in M^2, which is a Hilbert space with the inner product $\langle \phi, \psi \rangle = \phi^{0T}\psi^0 + \langle \phi^1, \psi^1 \rangle_2$. Let $S(t)$, $t \geqq 0$ be the C_0-semigroup corresponding to the free motion of (2.1), i.e.,

$$S(t)\phi = (x(t; \phi, 0), x_t(\phi, 0)), \qquad t \geqq 0, \quad \phi \in M^2.$$

With $B: \mathbb{R}^m \to M^2$ defined by

$$Bu = (B_0 u, 0), \qquad u \in \mathbb{R}^m,$$

the function

$$z(t; \phi, u) = S(t)\phi + \int_0^t S(t-s)Bu(s)\, ds, \qquad t \geqq 0, \tag{2.3}$$

is a mild solution of the abstract system

$$\dot{z}(t) = Az(t) + Bu(t), \qquad t \geqq 0, \quad z(0) = \phi \tag{Σ}$$

where the infinitesimal generator A of $S(\cdot)$ is given by

$$\operatorname{dom} A = \{\phi \in M^2 \mid \phi^1 \in W^{1,2}(-r, 0; \mathbb{R}^n),\ \phi^1(0) = \phi^0\},$$

$$A\phi = (A_0\phi^1(0) + A_1\phi^1(-r), \dot{\phi}^1).$$

The M^2-adjoint of A generates the M^2-adjoint semigroup $S(t)^*$, $t \geqq 0$:

$$\operatorname{dom} A^* = \{\phi \in M^2 \mid \phi^1 \in W^{1,2}(-r, 0; \mathbb{R}^n),\ \phi^1(-r) = A_1^T\phi^0\},$$

$$A^*\phi = (\phi^1(0) + A_0^T\phi^0, -\dot{\phi}^1).$$

The optimal control problem on a finite interval is: given $0 < T < \infty$ and $\phi \in M^2$ find the control $u \in L^2(0, T; \mathbb{R}^m)$ that minimizes the cost functional

$$\begin{aligned} J(u, \phi, T) &= \langle z(T; \phi, u), Gz(T; \phi, u) \rangle \\ &\quad + \int_0^T (\langle z(t; \phi, u), Wz(t; \phi, u) \rangle + u(t)^T Ru(t))\, dt \end{aligned} \tag{2.4}$$

where G, $W: M^2 \to M^2$ are defined by $G\phi = (G_0\phi^0, 0)$, $W\phi = (W_0\phi^0, 0)$ with G_0, W_0 being symmetric nonnegative matrices and $R = R^T$ positive definite. The optimal control $\bar{u}(\cdot)$ is given by the feedback law

$$\bar{u}(t) = -R^{-1}B^*\Pi(t)\bar{z}(t), \qquad 0 \leqq t \leqq T \tag{2.5}$$

where $\Pi(\cdot)$ is the unique, strongly continuous family of nonnegative self-adjoint and bounded operators satisfying the Riccati differential equation

$$\frac{d}{dt}\langle \psi, \Pi(t)\phi \rangle + \langle A\psi, \Pi(t)\phi \rangle + \langle \Pi(t)\psi, A\phi \rangle$$

$$-\langle \Pi(t)\psi, BR^{-1}B^*\Pi(t)\phi \rangle + \langle \psi, W\phi \rangle = 0 \quad \text{for } \phi, \psi \in \operatorname{dom} A, \qquad 0 \leqq t \leqq T,$$

$$\Pi(T) = G$$

and $\bar{z}(t)$ is the mild solution of

$$\dot{z}(t) = (A - BR^{-1}B^*\Pi(t))z(t), \qquad z(0) = \phi,$$

i.e.,

$$\bar{z}(t) = S(t)\phi - \int_0^t S(t-s)BR^{-1}B^*\Pi(s)\bar{z}(s)\, ds.$$

We also consider the infinite time horizon problem, that is the minimization of $J(u, \phi)$ given by (2.4) with $G \equiv 0$ and $T = \infty$, and assume that the system (2.1.1) is stabilizable, i.e., $A - BK$ generates an exponentially stable semigroup for some linear bounded operator $K : M^2 \to \mathbb{R}^m$. Then there exists a nonnegative, self-adjoint operator $\Pi \in \mathscr{L}(M^2)$ that maps dom A into dom A^* and satisfies the algebraic Riccati equation

$$A^*\Pi\phi + \Pi A\phi - \Pi BR^{-1}B^*\Pi\phi + W\phi = 0, \qquad \phi \in \text{dom } A. \tag{2.6}$$

If, in addition, (2.1.1) and W_0 have the property that any admissible control drives the state to zero, that is $J(u, \phi) < \infty$ implies $z(t; \phi, u) \to 0$, as $t \to \infty$, then Π is uniquely determined. This certainly is true if (2.1.1) with output W_0 is observable or if W_0 is simply nonsingular. Using the time-independent solution of (2.6) in the feedback law, (2.5) gives the optimal control and trajectory as in the finite time horizon case.

3. Finite-dimensional approximations. Our goal is to construct systems of ordinary differential equations, such that their solutions approximate the solution to the hereditary control problem in § 2. To this end, let Y^N, $N = 1, 2, \cdots$ be a sequence of finite-dimensional subspaces of $L^2(-r, 0; \mathbb{R}^n)$ with corresponding orthogonal projections p_1^N. Then $Z^N = \mathbb{R}^n \times Y^N$, $N = 1, 2, \cdots$ are finite-dimensional subspaces of M^2 with corresponding orthogonal projections $p^N\phi = (\phi^0, p_1^N\phi^1)$, $\phi \in M^2$. Suppose there is a sequence of linear operators $A^N : Z^N \to Z^N$ and let $S^N(t)$, $t \geqq 0$ be the uniformly continuous semigroups on M^2 generated by the bounded linear operators $A^Np^N : M^2 \to Z^N$, i.e.,

$$S^N(t)\phi = e^{A^Np^Nt}\phi, \qquad \phi \in M^2.$$

Remark. We extend A^N to all of M^2, because we want the semigroup $S^N(\cdot)$ acting on the whole space. Instead of letting the generator $A^Np^N = 0$ on $(Z^N)^\perp$, we could equally well choose another appropriate extension. All that is said about the control problems in Z^N and the corresponding semigroups $S^N(\cdot)$ in M^2 in this section remains valid, if we take the generator of $S^N(\cdot)$ to be $A^Np^N - \alpha(I - p^N)$ with some $\alpha \in \mathbb{R}$. For simplicity of exposition, we will make use of this possibility only at the end of the proof of Theorem 3.3 below.

Observing $B\xi = (B_0\xi, 0) \in Z^N$, $\xi \in \mathbb{R}^m$ for all N, we take the input operators $B^N : \mathbb{R}^m \to Z^N$ as $B^N\xi = B\xi$ to define finite-dimensional control systems on Z^N:

$$\dot{z}(t) = A^Nz(t) + B^Nu(t), \qquad t \geqq 0, \quad z(0) = p^N\phi, \tag{Σ^N}$$

where $u(\cdot) \in L^2_{\text{loc}}(0, \infty; \mathbb{R}^m)$ and $\phi \in M^2$. The optimal control $\bar{u}^N(\cdot)$ minimizing the functional $J^N(u, \phi, T)$, given by (2.4) with $z(\cdot; \phi, u)$ replaced by the solution $z^N(\cdot; \phi, u)$ of (Σ^N), is obtained as feedback by the Nth Riccati operator $\Pi^N(t)$ that satisfies the Riccati differential equation on Z^N with coefficients A^N, $(A^N)^*$, B^N, and W^N and terminal condition $\Pi^N(T) = G^N$ ($G^N = G|_{Z^N}$ and $W^N = W|_{Z^N}$). We write $\bar{z}^N(\cdot)$ for the corresponding optimal trajectory in Z^N.

In applications, the original system (Σ) is controlled by use of the approximate feedback instead of (2.5), i.e., $\bar{u}(t)$ is replaced by the so-called suboptimal control

$$\hat{u}^N(t) = -R^{-1}(B^N)^*\Pi^N(t)p^N\hat{z}^N(t), \qquad 0 \leqq t \leqq T \tag{3.1}$$

where $\hat{z}^N(t)$ is the mild solution of

$$\dot{z}(t) = (A - BR^{-1}(B^N)^*\Pi^N(t)p^N)z(t), \qquad z(0) = \phi.$$

From the uniform dissipativity assumption

(H1) There exists a constant $\omega \in \mathbb{R}$ such that

$$\langle A^N\phi, \phi\rangle \leqq \omega\|\phi\|^2 \quad \text{for all } \phi \in Z^N, \quad N = 1, 2, \cdots$$

follows the existence of a constant $M > 1$ such that

$$\|S^N(t)\phi\| \leqq M e^{\omega t}\|\phi\|, \qquad t \geqq 0, \quad \phi \in M^2. \tag{3.2}$$

Therefore, with

(H2) There exists a subset $D \subseteq \operatorname{dom} A$ and a real number $\lambda > \omega$, such that
(i) $(\lambda I - A)D$ is dense in M^2,
(ii) for all $\phi \in D$, $A^N p^N\phi \to A\phi$ as $N \to \infty$.

(H1), (H2) imply the strong convergence of $S^N(t)$ to $S(t)$ (Trotter–Kato Theorem [16, Chap. III, Thm. 4.5]). The same is true for $S^N{}^*(t)$ if (H2) with A, A^N replaced by A^*, $A^N{}^*$ holds (this will be denoted by (H2*)). Based on this and

(H3) $p^N\phi \to \phi$ for all $\phi \in M^2$

the following assertions were proved in [9] (also [13, Thm. 4.3]).

THEOREM 3.1. *Let* (H1)–(H3) *and* (H2*) *hold. Then, as* $N \to \infty$

(a) $\|\Pi^N(t)p^N - \Pi(t)\|_{\mathscr{L}(M^2)} \to 0.$

(b) $\hat{u}^N(t) \to \bar{u}(t), \quad \bar{u}^N(t) \to \bar{u}(t), \quad \hat{z}^N(t) \to \bar{z}(t), \quad \bar{z}^N(t) \to \bar{z}(t)$

the limits being uniform in t, $0 \leqq t \leqq T$.

(c) $J^N(\hat{u}^N, \phi, T) \to J(\bar{u}, \phi, T), \quad J^N(\bar{u}^N, \phi, T) \to J(\bar{u}, \phi, T).$

In the case of the infinite time horizon, we deal with the algebraic Riccati equations

$$(A^N)^*\Pi^N + \Pi^N A^N - \Pi^N B^N R^{-1}(B^N)^*\Pi^N + W^N = 0 \tag{3.3}$$

on Z^N, $N = 1, 2, \cdots$. If (A^N, B^N) is stabilizable, then there exists a nonnegative, self-adjoint solution Π^N of (3.3), governing the Nth optimal feedback. We will establish convergence of the Riccati operators Π^N using Gibson's arguments in [9] combined with the cross-product structure of the trajectories (see hypothesis (H4) below). This approach needs the assumption that the stabilizability of the hereditary system implies that the systems (A^N, B^N) are stabilizable in a uniform sense with respect to N (see (H5)). As to the investigation of stabilizability or exponential stability we will use a L^2-stability criterion due to Datko [7]. We state here a special version (a proof may also be found in [19]).

LEMMA 3.2. *Let* $S(t)$, $t \geqq 0$ *be a* C_0*-semigroup of bounded linear operators in a Banach space* X *satisfying*

$$\|S(t)\|_{\mathscr{L}(X)} \leqq M_1 e^{\alpha_1 t}, \qquad t \geqq 0$$

and

$$\int_0^\infty \|S(t)x\|_X^2\, dt \leqq c_1\|x\|_X^2, \qquad x \in X \tag{3.4}$$

for some constants M_1, α_1, $c_1 > 0$. *Then there exists an exponent* $\alpha = \alpha(c_1, M_1, \alpha_1) > 0$ *and a constant* $M = M(c_1, M_1, \alpha_1) > 0$ *such that*

$$\|S(t)\|_{\mathcal{L}(X)} \leqq M e^{-\alpha t}, \qquad t \geqq 0. \tag{3.5}$$

Note that if we can prove (3.4) for the semigroups $S^N(\cdot)$ with c_1 independent on N, then, by (3.2), (3.5) yields the exponential stability of $S^N(\cdot)$ uniformly with respect to N. Moreover, observe that the estimate (3.4) is equivalent to

$$\int_0^\infty |(S(t)x)^0|^2_{\mathbb{R}^n}\, dt \leqq c_2 \|x\|^2_{M^2}, \qquad x \in M^2, \tag{3.6}$$

for some constant $c_2 > 0$, if $S(\cdot)$ is the solution semigroup on M^2 as defined in § 2. By Fubini's Theorem, this equivalence follows directly from the state concept (2.2) and is a special feature of the semigroup associated with the retarded functional differential equation.

As far as we want the semigroups $S^N(\cdot)$ to be suitable approximations for $S(\cdot)$, it seems to be natural to demand the equivalence of (3.4) and (3.6) also with regard to $S^N(\cdot)$. We call this the VDP-property of the approximations (meaning the vector-component dominance is preserved). It plays an essential role in our approach to stability questions in context with the infinite time horizon control problem.

Suppose for N sufficiently large there exists a solution Π^N to the Nth algebraic Riccati equation (3.3). Then with

$$\bar{A}^N = A^N - B^N R^{-1} (B^N)^* \Pi^N,$$

$\bar{A}^N p^N$ generates an exponentially stable semigroup $\bar{S}^N(\cdot)$ on M^2. We introduce the projection $V: M^2 \to \mathbb{R}^n$

$$V(\phi^0, \phi^1) = \phi^0$$

and want the following hypothesis to be valid.

(H4) Provided N is sufficiently large, there exists a $c_1 \geqq 0$, independent on N, such that for all $\phi \in Z^N$

$$\int_0^\infty \|\bar{S}^N(t)\phi\|^2_{M^2}\, dt \leqq c_1 \|\phi\|^2_{M^2}$$

if and only if there exists a $c_2 \geqq 0$, independent on N, such that for all $\phi \in Z^N$

$$\int_0^\infty |V\bar{S}^N(t)\phi|^2_{\mathbb{R}^n}\, dt \leqq c_2 \|\phi\|^2_{M^2}.$$

If Π is a nonnegative self-adjoint solution to the algebraic Riccati equation (2.6) for the hereditary control problem, define the operators $\tilde{A}^N : Z^N \to Z^N$, $N = 1, 2, \cdots$ by

$$\tilde{A}^N = A^N - BR^{-1}B^*\Pi$$

and let $\tilde{S}^N(\cdot)$ be the uniformly continuous semigroup on M^2 generated by $\tilde{A}^N p^N$, i.e., $\tilde{S}^N(t) = e^{\tilde{A}^N p^N t}$, $t \geqq 0$.

Intending to provide the existence and uniform boundedness of the operators Π^N, we demand the following.

(H5) If the hereditary system (Σ) is stabilizable, then there exist constants $M, \beta > 0$ such that for N sufficiently large

$$\|\tilde{S}^N(t)|_{Z^N}\| \leqq M e^{-\beta t}, \qquad t \geqq 0.$$

THEOREM 3.3. *Let* (H1)-(H5), (H2*) *hold and assume* W_0 *is nonsingular. If the hereditary system is stabilizable, then*

(a) *For* N *sufficiently large there exists a solution* Π^N *to the* N*th algebraic Riccati equation* (3.3) *and*

$$\|\Pi^N p^N - \Pi\|_{\mathscr{L}(M^2)} \to 0, \qquad N \to \infty.$$

(b) *The optimal and suboptimal controls and trajectories and the corresponding costs converge as in Theorem* 3.1(b),(c).

Proof. As in Theorem 7.4 of [9], we first consider the Nth problem (Σ^N) with initial value $\phi \in M^2$, when it is controlled by the feedback $\tilde{u}^N(t) = -R^{-1}B^*\Pi p^N z(t)$. The evolution of the state in time is then described by $\tilde{z}^N(t) = \tilde{S}^N(t)p^N\phi$ and the corresponding costs can be estimated using (H5):

$$J^N(\tilde{u}^N, \phi) = \int_0^\infty (\langle \tilde{z}^N(t), W\tilde{z}^N(t)\rangle + \tilde{u}^N(t)^T R\tilde{u}^N(t))\, dt$$

$$\leqq \frac{M}{2\beta}(|W_0| + |R^{-1}||B_0|^2\|\Pi\|^2)\|\phi\|^2.$$

Thus, there exists a nonnegative self-adjoint solution Π^N of the Nth algebraic Riccati equation for N sufficiently large and

$$\langle p^N\phi, \Pi^N p^N\phi\rangle = J^N(\tilde{u}^N, \phi) \leqq J^N(\tilde{u}^N, \phi) \leqq c_1\|\phi\|^2$$

with some constant c_1, which does not depend on N. Therefore, there is an index N_0 such that

$$\|\Pi^N p^N\| \leqq c_1, \qquad N \geqq N_0. \tag{3.7}$$

The convergence statement in (a) now follows from Theorem 6.9 of [9], once we have shown

$$\|\bar{S}^N(t)\| \leqq \bar{M}e^{-\alpha t}, \qquad t \geqq 0, \quad N \geqq N_0 \tag{3.8}$$

for some constants $\bar{M}$, $\alpha > 0$.

Since $W_0 > 0$ we have $|\xi|^2 \leqq \mu^{-1}\xi^T W_0 \xi$, $\xi \in \mathbb{R}^n$, where μ is the minimum eigenvalue of W_0. Following the arguments given in [9] (proof of Theorem 7.5) let $\phi \in Z^N$ and define $\bar{z}^N(t) = \bar{S}^N(t)\phi = (x^N(t), y^N(t))$ with $x^N(t) \in \mathbb{R}^n$, $y^N(t) \in Y^N$, $t \geqq 0$. Furthermore, note that $\langle R^{-1}B^*\Pi^N\bar{z}^N(t), B^*\Pi^N\bar{z}^N(t)\rangle_{\mathbb{R}^m} \geqq 0$, so that

$$\int_0^\infty |x^N(t)|^2\, dt \leqq \mu^{-1}\int_0^\infty x^N(t)^T W_0 x^N(t)\, dt$$

$$\leqq \mu^{-1}\int_0^\infty \langle \bar{z}^N(t), (W + \Pi^N BR^{-1}B^*\Pi^N)\bar{z}^N(t)\rangle_{M^2}\, dt$$

$$= \mu^{-1}\left\langle \phi, \int_0^\infty (\bar{S}^N)^*(t)(W + \Pi^N BR^{-1}B^*\Pi^N)\bar{S}^N(t)\phi\, dt \right\rangle$$

$$= \mu^{-1}\langle \phi, \Pi^N\phi\rangle,$$

applying Corollary 4.2 of [9] to the semigroups $S^N(\cdot)$. It follows that $\int_0^\infty |V\bar{S}^N(t)\phi|^2\,dt \leqq \mu^{-1}c_1\|\phi\|^2$, $N \geqq N_0$ and, by hypothesis (H4), $\int_0^\infty \|\bar{S}^N(t)\phi\|^2\,dt \leqq c_2\|\phi\|^2$ with some $c_2 > 0$. Also, because $\bar{S}^N$ is generated by $(A^N - BR^{-1}B^*\Pi^N)p^N$, (3.2) and (3.7) imply (e.g., [16, Chap. III, Thm. 1.1]) the existence of constants M_1, α_1 such that $\|\bar{S}^N(t)\| \leqq M_1 e^{\alpha_1 t}$, $t \geqq 0$, $N \geqq N_0$. Now Lemma 3.2 assures that there are constants M, $\alpha > 0$ such that

$$\|\bar{S}^N(t)|_{Z^N}\| \leqq M e^{-\alpha t}, \qquad t \geqq 0, \quad N \geqq N_0.$$

But this proves (3.8), since, as we mentioned in the remark above, we may replace $A^N p^N$ by $A^N p^N - \alpha(I - p^N)$, so that $\|\bar{S}^N(t)|_{(Z^N)^\perp}\| \leqq e^{-\alpha t}$, while the finite-dimensional control problems on Z^N remain completely unchanged. Statement (b) follows from (a) in a manner similar to that in the finite time horizon case (see [9]). □

4. The piecewise linear approximation scheme. This section presents a special approximation scheme using so-called piecewise linear functions. We prove via several lemmas that this scheme satisfies the hypotheses (H1)-(H3) and (H2*). Then we show that it has the VDP-property, so that (H4) is valid. Furthermore, we investigate the characteristic matrix of the approximate systems to establish results on the eigenvalue behavior when the approximation index increases. This enables us to conclude (H5). Finally, after remarks on the treatment of multiple and distributed delays, we present some of our numerical findings.

4.1. Projection onto spaces of piecewise linear functions. For $N = 1, 2, \cdots$ we subdivide the interval $[-r, 0]$ into the subintervals $I_j^N = [t_j^N, t_{j-1}^N)$, $j = 2, \cdots, N$ and $I_1^N = [t_1^N, 0]$ by defining the meshpoints

$$t_j^N = -\frac{jr}{N}, \quad j = 0, \cdots, N.$$

For each $N \in \mathbb{N}$ the set Y^N of all functions $[-r, 0] \to \mathbb{R}^n$ that are polynomials of degree one on every interval I_j^N is commonly called a space of piecewise linear functions on $[-r, 0]$. A basis of Y^N is given, written in a simplified notation, by the $2N$ matrix functions

$$\begin{aligned} e_{2j-1}^N(s) &= \chi_j^N(s) \cdot I, \\ e_{2j}^N(s) &= \left(\frac{2N}{r} + 2j - 1\right)\chi_j^N(s) \cdot I, \end{aligned} \qquad j = 1, \cdots, N$$

where I denotes the $n \times n$ identity matrix and χ_j^N is the characteristic function of I_j^N. The pairs $\hat{e}_0 = (I, 0)$ and $\hat{e}_j^N = (0, e_j^N)$ are an orthogonal basis of the $n(2N+1)$-dimensional product space $Z^N = \mathbb{R}^n \times Y^N$, $N = 1, 2, \cdots$. The orthogonal projections $p^N : M^2 \to Z^N$ are of the form $p^N(\phi^0, \phi^1) = (\phi^0, p_1^N \phi^1)$, where p_1^N is the L^2-orthogonal projection from L^2 onto Y^N. With

$$Q^N = (\langle \hat{e}_j^N, \hat{e}_k^N \rangle) = \operatorname{diag}\left(1, \frac{r}{N}, \frac{r}{3N}, \cdots, \frac{r}{N}, \frac{r}{3N}\right) \otimes I \tag{4.1}$$

we have

$$\begin{aligned} &\langle \phi, \psi \rangle = \alpha^N(\phi)^T Q^N \alpha^N(\psi), \qquad \phi, \psi \in Z^N, \\ &\alpha^N(p^N\phi) = (Q^N)^{-1} \operatorname{col}\,(\phi^0, \langle e_1^N, \phi^1 \rangle_2, \cdots, \langle e_{2N}^N, \phi^1 \rangle_2), \qquad \phi \in M^2 \end{aligned} \tag{4.2}$$

where the components of the coefficient vector $\alpha^N(p^N\phi)$ are given by $\alpha_0 = \phi^0$ and

$$\begin{aligned} \alpha^N_{2j-1}(p^N\phi) &= \frac{N}{r}\int_{I^N_j} \phi^1(s)\,ds, \\ \alpha^N_{2j}(p^N\phi) &= \frac{3N}{r}\int_{I^N_j} e^N_{2j}(s)\phi^1(s)\,ds, \end{aligned} \qquad j = 1, \cdots, N. \tag{4.3}$$

We frequently will use the abbreviations $\phi^N_j = \alpha^N_j(p^N\phi)$ and $\phi^N = p^N_1\phi^1$. Note that

$$\begin{aligned} |\phi^N_{2j-1}| &\leqq \left(\frac{N}{r}\right)^{1/2} \|\phi^1\|_2 \\ |\phi^N_{2j}| &\leqq \left(\frac{3N}{r}\right)^{1/2} \|\phi^1\|_2, \end{aligned} \qquad j = 1, \cdots, N. \tag{4.4}$$

Obviously, the spaces Z^N are not contained in the domain of the generator A of the hereditary semigroup, since the elements of Y^N are not differentiable on $[-r, 0]$. Nevertheless, the action of A and A^* can be approximated by operators on Z^N imitating, heuristically speaking, the delta distributions in the derivatives of discontinuous functions by operators δ^N in Y^N. Following the ideas developed in [13], we need the operators δ^N for each point, where the piecewise linear functions in Y^N may have jumps.

We define $\delta^{N-}_j, \delta^{N+}_j : \mathbb{R}^n \to Y^N$, $N = 1, 2, \cdots$ by

$$\delta^{N-}_j(\xi) = \left(\frac{N}{r}e^N_{2j+1} + \frac{3N}{r}e^N_{2j+2}\right)\xi, \qquad j = 0, \cdots, N-1$$

and

$$\delta^{N+}_j(\xi) = \left(\frac{N}{r}e^N_{2j-1} - \frac{3N}{r}e^N_{2j}\right)\xi, \qquad j = 1, \cdots, N.$$

PROPOSITION 4.1. (a) *For any $\xi \in \mathbb{R}^n$ and $\phi^1 \in L^2(-h, 0; \mathbb{R}^n)$*

$$\langle \delta^{N-}_j(\xi), \phi^1\rangle_2 = \xi^T\phi^N(t^{N-}_j), \qquad j = 0, \cdots, N-1, \tag{4.5}$$

$$\langle \delta^{N+}_j(\xi), \phi^1\rangle_2 = \xi^T\phi^N(t^N_j), \qquad j = 1, \cdots, N. \tag{4.6}$$

In (4.5) *and throughout the paper, we use the notation $\phi^N(t^{N-}_j)$ for the left side limit of ϕ^N at t^N_j.*

(b) *The norms of the δ-operators are*

$$\|\delta^{N-}_j\| = 2\left(\frac{N}{r}\right)^{1/2}, \qquad j = 0, \cdots, N-1,$$

$$\|\delta^{N+}_j\| = 2\left(\frac{N}{r}\right)^{1/2}, \qquad j = 1, \cdots, N.$$

Proof. (a) Let $1 \leqq j \leqq N-1$. Using (4.3) and observing that

$$e^N_{2j-1}(t^{N-}_{j-1}) = e^N_{2j}(t^{N-}_{j-1}) = I,$$

we get

$$\begin{aligned} \langle \delta^{N-}_j(\xi), \phi^1\rangle_2 &= \int_{-r}^0 \xi^T\left(\frac{N}{r}e^N_{2j+1} + \frac{3N}{r}e^N_{2j+2}\right)(s)\phi^1(s)\,ds \\ &= \xi^T(\phi^N_{2j+1} + \phi^N_{2j+2}) = \xi^T\phi^N(t^{N-}_j). \end{aligned}$$

The proof for δ_0^{N-} is analogous. The statement on δ_j^{N+} follows similarly from the fact that $e_{2j-1}^N(t_j^N) = I$ and $e_{2j}^N(t_j^N) = -I$, $j = 1, \cdots, N$.

(b) Since the elements e_j^N are orthogonal, we have

$$\|\delta_j^{N+}(\xi)\|_2^2 = |\xi|^2\left(\left(\frac{N}{r}\right)^2\|e_{2j-1}^N\|^2 + \left(\frac{3N}{r}\right)^2\|e_{2j}^N\|^2\right)$$

$$= |\xi|^2\left(\frac{N}{r} + \frac{3N}{r}\right) \quad \text{by (4.1).}$$

The proof for δ_j^{N-} is analogous. □

Next we give convergence estimates for the piecewise linear projections of sufficiently smooth functions.

LEMMA 4.2. *For* $\phi^1 \in W^{2,\infty}(-r, 0; \mathbb{R}^n)$

(a) $$\|\phi^N - \phi^1\|_\infty \leqq \frac{c_1}{N^2}\|\ddot{\phi}^1\|_\infty, \qquad N = 1, 2, \cdots.$$

(b) $$\|D^+\phi^N - D\phi^1\|_\infty \leqq \frac{c_2}{N}\|\ddot{\phi}^1\|_\infty,$$

The constants c_1, c_2 *do not depend on* N *or* ϕ^1.

Proof. Applying the Peano Kernel Theorem (see, for instance, Theorem 1.3 of [20]) to the functionals $F_S, G_S : W^{2,\infty}(I_j^N) \to \mathbb{R}^n$, $s \in I_j^N$, given by $F_S(\phi^1) = \phi^1(s) - \phi^N(s)$ and $G_s(\phi^1) = D\phi^1(s) - D\phi^N(s)$, we obtain estimates involving integrals of $\ddot{\phi}^1$ and their projections onto Y^N. These (respectively, their derivatives) are estimated by expansion according to the basis $\{e_j^N\}$ and then using (4.4) and $|e_j^N(s)| \leqq 1$. □

As an immediate consequence from Lemma 4.2(a), we see that the subspaces Z^N defined in this section satisfy hypothesis (H3), since the set $\{(\phi^0, \phi^1) \in M^2 | \phi^0 \in \mathbb{R}^n, \phi^1 \in W^{2,\infty}(-r, 0; \mathbb{R}^n)\}$ is dense in M^2 and $\|p^N\| = 1$ for all N.

4.2. The approximating semigroups and their generators.

DEFINITION 4.3. For $\phi = (\phi^0, \phi^1) \in Z^N$ we define

$$A^N(\phi^0, \phi^1) = \Big(A_0\phi^0 + A_1\phi^1(-r),\ D^+\phi^1 + \delta_0^{N-}(\phi^0 - \phi^1(0)) + \sum_{j=1}^{N-1}\delta_j^{N-}(\phi^1(t_j^N) - \phi^1(t_j^{N-}))\Big).$$

Since $D^+\phi^1 \in Y^N$ for $\phi^1 \in Y^N$ it is clear that A^N is a linear operator $Z^N \to Z^N$.

LEMMA 4.4. *The adjoint of* A^N *is given by*

$$(A^N)^*(\psi^0, \psi^1) = \Big(A_0^T\psi^0 + \psi^1(0),\ -D^+\psi^1 + \sum_{j=1}^{N-1}\delta_j^{N+}(\psi^1(t_j^{N-}) - \psi^1(t_j^N)) + \delta_N^{N+}(A_1^T\psi^0 - \psi^1(-r))\Big)$$

for $(\psi^0, \psi^1) \in Z^N$.

Proof. For $\phi, \psi \in Z^N$, integration by parts of the term in $\langle\psi, A^N\phi\rangle$ involving $D^+\phi^1$ yields

$$\langle\psi^1, D^+\phi^1\rangle_2 = \sum_{j=0}^{N-1}\psi^1(t_j^{N-})^T\phi^1(t_j^{N-}) - \sum_{j=1}^{N}\psi^1(t_j^N)^T\phi^1(t_j^N) - \langle D^+\psi^1, \phi^1\rangle_2.$$

Furthermore, by Proposition 4.1

$$\langle\psi^1, \delta_0^{N-}(\phi^0 - \phi^1(0))\rangle_2 = (\phi^0 - \phi^1(0))^T\psi^1(0^-) = \psi^1(0)^T\phi^0 - \psi^1(0)^T\phi^1(0),$$

and

$$\psi^{0T} A_1 \phi^1(-r) = \langle \delta_N^{N+}(A_1^T \psi^0), \phi^1 \rangle_2,$$

so that

$$\langle (A^N)^*(\psi^0, \psi^1), (\phi^0, \phi^1) \rangle = (A_0^T \psi^0 + \psi^1(0))^T \phi^0 + \langle \delta_N^{N+}(A_1^T \psi^0) - D^+ \psi^1, \phi^1 \rangle_2 + \Delta$$

where Δ is given by

$$\Delta = -\psi^1(0)^T \phi^1(0) + \sum_{j=0}^{N-1} \psi^1(t_j^{N-})^T \phi^1(t_j^{N-}) - \sum_{j=1}^{N} \psi^1(t_j^N)^T \phi^1(t_j^N)$$

$$+ \sum_{j=1}^{N-1} \langle \psi^1, \delta_j^{N-}(\phi^1(t_j^N) - \phi^1(t_j^{N-})) \rangle_2.$$

Here the last sum is transformed using (4.5). By the continuity of ϕ^1, $\psi^1 \in Y^N$ at 0, the first two terms can also be summed up. This yields

$$\Delta = \sum_{j=1}^{N-1} (\psi^1(t_j^{N-}) - \psi^1(t_j^N))^T \phi^1(t_j^N) - \psi^1(-r)^T \phi^1(-r).$$

The result follows by applications of (4.6). □

The next lemma shows that the operators A^N satisfy the uniform dissipativity condition (H1).

LEMMA 4.5. *For all N and all* $\phi \in Z^N$, $\langle A^N \phi, \phi \rangle \leqq \omega \|\phi\|^2$, *with* ω *being the dissipativity constant of A, i.e.,* $\omega = \frac{1}{2} + |A_0| + \frac{1}{2}|A_1|^2$.

Proof. From the definition of A^N and Proposition 4.1, it follows that

$$\langle A^N \phi, \phi \rangle \leqq (A_0 \phi^0)^T \phi^0 + |A_1||\phi^1(-r)||\phi^0| + \sum_{j=1}^{N} \int_{I_j^N} (D^+ \phi^1(s))^T \phi^1(s)\, ds$$

$$+ (\phi^0 - \phi^1(0))^T \phi^1(0^-) + \sum_{j=1}^{N-1} (\phi^1(t_j^N) - \phi^1(t_j^{N-}))^T \phi^1(t_j^{N-})$$

for $\phi \in Z^N$. Using

$$\int_{I_j^N} (D^+ \phi^1(s))^T \phi^1(s)\, ds = \frac{1}{2}(|\phi^1(t_{j-1}^{N-})|^2 - |\phi^1(t_j^N)|^2)$$

and the inequality $\xi^T \eta \leqq \frac{1}{2}(|\xi|^2 + |\eta|^2)$, $\xi, \eta \in \mathbb{R}^n$, we get

$$\langle A^N \phi, \phi \rangle \leqq (A_0 \phi^0)^T \phi^0 + \frac{1}{2}(|A_1|^2 \|\phi\|^2 + |\phi^1(-r)|^2) + \frac{1}{2} \sum_{j=0}^{N-1} |\phi^1(t_j^{N-})|^2$$

$$- \frac{1}{2} \sum_{j=1}^{N} |\phi^1(t_j^N)|^2 + |\phi^0||\phi^1(0)| - |\phi^1(0)|^2$$

$$+ \frac{1}{2} \sum_{j=1}^{N-1} |\phi^1(t_j^N)|^2 - \frac{1}{2} \sum_{j=1}^{N-1} |\phi^1(t_j^{N-})|^2$$

$$\leqq (A_0 \phi^0)^T \phi^0 + \frac{1}{2}|\phi^0|^2 + \frac{1}{2}|A_1|^2 \|\phi\|^2. \qquad \square$$

Looking for appropriate sets to be used in (H2) and (H2*), we define

$$D = \{(\phi^1(0), \phi^1) \in M^2 \,|\, \phi^1 \in W^{2,\infty}(-r, 0; \mathbb{R}^n)\}$$

and

$$D^* = \{(\psi^0, \psi^1) \in \operatorname{dom} A^* \mid \psi^1 \in W^{2,\infty}(-r, 0; \mathbb{R}^n)\}.$$

The following lemma establishes (H2)(ii) and (H2*)(ii).

LEMMA 4.6. (a) *There is a constant $c > 0$, such that for all N and all $\phi \in D$*

$$\|A^N p^N \phi - A\phi\| \leqq \frac{c}{N} \|\ddot{\phi}^1\|_\infty.$$

(b) *For all $\psi \in D^*$*

$$\|(A^N)^* p^N \psi - A^* \psi\| \to 0 \quad \text{as } N \to \infty.$$

Proof. (a) From the definition of A and A^N we have the estimate

$$\begin{aligned}\|A^N p^N \phi - A\phi\| \leqq |A_1||\phi^N(-r) - \phi^1(-r)| + \|D^+\phi^N - D\phi^1\|_2 \\ + \|\delta_0^{N-}(\phi^0 - \phi^N(0))\|_2 \\ + \left\| \sum_{j=1}^{N-1} \delta_j^{N-}(\phi^N(t_j^N) - \phi^N(t_j^{N-})) \right\|_2.\end{aligned}$$

By Proposition 4.1 and Lemma 4.2 for $\phi \in D$ this yields

$$\|A^N p^N \phi - A\phi\| \leqq \left(\frac{c_1}{N^2} + \frac{c_2}{N} + 2\left(\frac{N}{r}\right)^{1/2} \frac{c_3}{N^2}\right) \|\ddot{\phi}^1\|_\infty + \|\Delta^N(\phi^1)\|_2$$

with some constants c_1, c_2, c_3. Taking advantage of the orthogonality of $e_1^N, \cdots, e_{2N}^N$ in L^2, we have for $\xi \in \mathbb{R}^n$

$$\left\| \sum_{j=1}^{N-1} \delta_j^{N-}(\xi) \right\|_2^2 \leqq 4(N-1)\frac{N}{r}|\xi|^2.$$

Hence,

$$\|\Delta^N(\phi^1)\|_2 \leqq 2\left((N-1)\frac{N}{r}\right)^{1/2} \max_{j=1,\cdots,N} |\phi^N(t_j^N) - \phi^N(t_j^{N-})| \leqq \frac{\text{const.}}{N} \|\ddot{\phi}^1\|_\infty.$$

This proves (a). Item (b) follows using similar estimates on the δ_j^{N+} terms and by the fact that $A_1^T \psi^0 = \psi^1(-r)$ for $\psi \in D^*$. □

In order to apply Theorem 3.1 to A^N and $(A^N)^*$ of this section, it remains to show (H2)(i) and (H2*)(i).

LEMMA 4.7. *The sets $(\lambda - A)D$ and $(\lambda - A^*)D^*$ are dense in M^2, if $\lambda > \omega$ as given in Lemma* 4.5.

Proof. We know from semigroup theory that $\{\lambda \in \mathbb{C} \mid \operatorname{Re} \lambda > \omega\}$ is contained in the resolvent sets $\rho(A)$ and $\rho(A^*)$. Hence, given $\psi \in \mathbb{R}^n \times C^1(-h, 0; \mathbb{R}^n)$ the equation $(\lambda - A)\phi = \psi$ has a unique solution $\phi = (\phi^1(0), \phi^1) \in \operatorname{dom} A$, which by the definition of A satisfies $\lambda\phi^1 - \dot{\phi}^1 = \psi^1$. But this implies $\dot{\phi}^1 = \lambda\phi^1 - \psi^1$ is continuous and differentiable. In fact, $\ddot{\phi}^1 = \lambda\dot{\phi}^1 - \dot{\psi}^1$ is continuous, so that $\phi \in \mathbb{R}^n \times C^2(-h, 0; \mathbb{R}^n)$. Hence the dense set $\mathbb{R}^n \times C^1$ is contained in $(\lambda - A)(\mathbb{R}^n \times C^2 \cap \operatorname{dom} A)$, which is a subset of $(\lambda - A)D$. The same arguments can be applied to $(\lambda - A^*)D^*$. □

Summarizing, we can now say that the semigroups $S^N(\cdot)$ and $S^N(\cdot)^*$ generated by $A^N p^N$ and $(A^N)^* p^N$ strongly converge to the hereditary semigroups $S(\cdot)$ and $S(\cdot)^*$ so that for $T < \infty$ we can approximate the Riccati operators and the optimal controls by solving finite-dimensional problems in Z^N.

4.3. Vector dominance preservation. With respect to the basis $\hat{e}_0, \hat{e}_1^N, \cdots, \hat{e}_{2N}^N$ of Z^N and the canonical basis of $\mathbb{R}^n$ the matrices $[A^N]$ and $[A^N*]$ representing A^N and A^N* are given by

$$[A^N]=(Q^N)^{-1}H^N \quad \text{and} \quad [A^N*]=(Q^N)^{-1}H^{NT}$$

with

$$H^N=\begin{bmatrix} \langle \hat{e}_0, A^N\hat{e}_0\rangle & \cdots & \langle \hat{e}_0, A^N\hat{e}_{2N}^N\rangle \\ \vdots & & \vdots \\ \langle \hat{e}_{2N}^N, A^N\hat{e}_0\rangle & \cdots & \langle \hat{e}_{2N}^N, A\hat{e}_{2N}^N\rangle \end{bmatrix}, \qquad N=1,2,\cdots.$$

For the computation of the entries of H^N, observe that the derivatives of the basis elements are given by

$$D^+e_{2j-1}^N=0 \quad \text{and} \quad D^+e_{2j}^N=\frac{2N}{r}e_{2j-1}^N, \quad j=1,\cdots,N.$$

The inner products in H^N are evaluated using the orthogonality of the basis elements and Proposition 4.1. The result is the $n(2N+1)$ square matrix

$$H^N=\begin{bmatrix} A_0 & 0 & \cdots & 0 & A_1 & -A_1 \\ k_0 & h & & & & \\ & k & & 0 & & \\ & & \ddots & & & \\ & 0 & & k & & h \end{bmatrix} \tag{4.7}$$

where $k_0=\begin{pmatrix} I \\ I \end{pmatrix}$ is a $2n\times n$ matrix and

$$h=\begin{pmatrix} -I & I \\ -I & -I \end{pmatrix}, \qquad k=\begin{pmatrix} I & -I \\ I & -I \end{pmatrix}$$

are $2n\times 2n$ matrices. Numerical algorithms solving high-order systems with coefficients $[A^N]$ might take considerable advantage of the fact that H^N has band structure (not the case for the Legendre methods [10]-[12]) and that Q^N is diagonal (not the case for spline methods [13]).

In the following, we exploit the structure of the matrix $[A^N]$ to deduce the VDP property and (H4) for our piecewise linear approximations.

LEMMA 4.8. *If there is a $c_2>0$ independent on N such that for all $\phi\in Z^N$*

$$\int_0^\infty |VS^N(t)\phi|^2\,dt\leqq c_2\|\phi\|^2,$$

then

$$\int_0^\infty \|S^N(t)\phi\|^2\,dt\leqq c_1\|\phi\|^2, \qquad \phi\in Z^N,$$

for some $c_1\geqq 0$ not depending on N or ϕ.

Proof. Let $\phi\in Z^N$ and set $S^N(t)\phi=e^{A^{N_p N_t}}\phi=e^{A^N t}\phi=(w_0^N(t), w_1^N(t))\in Z^N$, $t\geqq 0$. With $w_1^N(t)=\sum_{j=1}^{2N} e_j^N w_j^N(t)$ the coefficient vector $\operatorname{col}(w_0^N(t), w_1^N(t), \cdots, w_{2N}^N(t))$ is the solution of

$$\frac{d}{dt}\begin{bmatrix} w_0^N(t) \\ w_1^N(t) \\ \vdots \\ w_{2N}^N(t) \end{bmatrix}=[A^N]\begin{bmatrix} w_0^N(t) \\ w_1^N(t) \\ \vdots \\ w_{2N}^N(t) \end{bmatrix}, \qquad t\geqq 0, \qquad \begin{matrix} w_0^N(0)=\phi^0, \\ w_j^N(0)=\phi_j^N. \end{matrix} \tag{4.8}$$

A view of the rows of $[A^N]$ reveals that (4.8) implies

$$\dot{v}_1^N(t) = -\frac{N}{r} a v_1^N(t) + \frac{N}{r} v_0^N(t),$$

$$\dot{v}_j^N(t) = -\frac{N}{r} a v_j^N(t) + \frac{N}{r} b v_{j-1}^N(t), \qquad j = 2, \cdots, N, \tag{4.9}$$

$$v_j^N(0) = \operatorname{col}(\phi_{2j-1}^N, \phi_{2j}^N), \qquad j = 1, \cdots, N$$

where $v_j^N(t) = \operatorname{col}(w_{2j-1}^N(t), w_{2j}^N(t)) \in \mathbb{R}^{2n}$, $v_0^N(t) = \operatorname{col}(w_0^N(t), 3w_0^N(t))$ and

$$a = \begin{pmatrix} 1 & -1 \\ 3 & 3 \end{pmatrix} \otimes I, \qquad b = \begin{pmatrix} 1 & -1 \\ 3 & -3 \end{pmatrix} \otimes I.$$

Using (4.4), we get

$$|v_j^N(0)|^2 \leqq \frac{4N}{r} \|\phi\|^2, \qquad j = 1, \cdots, N. \tag{4.10}$$

To estimate the solutions of (4.9), we make use of the fact that, if $f \in L^2(0, \infty; \mathbb{R})$, $\alpha > 0$ and $g(t) = \int_0^t e^{-\alpha(t-s)} f(s)\, ds$, $t \geqq 0$, then $g \in L^2(0, \infty; \mathbb{R})$ and $\int_0^\infty g^2(t)\, dt \leqq 1/\alpha^2 \int_0^\infty f^2(t)\, dt$ (cf. [9, Lemma 7.3]). The first equation in (4.9) yields

$$v_1^N(t) = e^{-(N/r)at} v_1^N(0) + \int_0^t e^{-(N/r)a(t-s)} \frac{N}{r} v_0^N(s)\, ds, \qquad t \geqq 0.$$

With $\tilde{v}_1^N(t) = v_1^N(t) - e^{-(N/r)at} v_1^N(0)$, $t \geqq 0$, we get

$$\begin{aligned} \int_0^\infty |\tilde{v}_1^N(t)|^2\, dt &\leqq \frac{r^2}{N^2} \text{const.} \int_0^\infty \frac{N^2}{r^2} |v_0^N(t)|^2\, dt \\ &\leqq \text{const.} \int_0^\infty |w_0^N(t)|^2\, dt \\ &= \text{const.} \int_0^\infty |VS^N(t)\phi|^2\, dt \\ &\leqq \text{const.} \|\phi\|^2, \end{aligned}$$

by assumption, where the constants do not depend on N. Using (4.10), it follows that

$$\int_0^\infty |v_1^N(t)|^2\, dt \leqq \text{const.} \|\phi\|^2. \tag{4.11}$$

Estimating the solutions of the other equations in (4.9) by the same method, we obtain

$$\int_0^\infty |v_j^N(t)|^2\, dt \leqq \text{const.} \|\phi\|^2 \quad \text{if} \quad \int_0^\infty |v_{j-1}^N(t)|^2\, dt \leqq \text{const.} \|\phi\|^2, \quad j = 2, \cdots, N.$$

So, from (4.11), by induction we get

$$\int_0^\infty |v_j^N(t)|^2\, dt \leqq \text{const.} \|\phi\|^2, \qquad j = 1, \cdots, N.$$

But this proves the assertion of the lemma, because by (4.2)

$$\|S^N(t)\phi\|^2 = |w_0^N(t)|^2 + \sum_{j=1}^{N} \left[\frac{r}{N} |w_{2j-1}^N(t)|^2 + \frac{r}{3N} |w_{2j}^N(t)|^2 \right]$$

$$= |w_0^N(t)|^2 + \sum_{j=1}^{N} \frac{r}{N} |dv_j^N(t)|^2$$

where

$$d = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1/3} \end{pmatrix} \otimes I$$

so that

$$\int_0^\infty \|S^N(t)\phi\|^2 \, dt \leqq \int_0^\infty |w_0^N(t)|^2 \, dt + \frac{r}{N} \sum_{j=1}^{N} |d|^2 \int_0^\infty |v_j^N(t)|^2 \, dt$$

$$\leqq \text{const.} \, \|\phi\|^2. \qquad \square$$

COROLLARY 4.9. *The semigroups* $\bar{S}^N(\cdot)$ *generated by* $\bar{A}^N p^N = (A^N - B^N R^{-1} (B^N)^* \Pi^N) p^N$ *satisfy hypothesis* (H4).

Proof. Since $B^N : \mathbb{R}^n \to Z^N$ is represented by the $n(2N+1) \times m$ matrix $[B^N] = \text{col}\,(B_0, 0, \cdots, 0)$, the matrix $[\bar{A}^N]$ differs from $[A^N]$ only in the first n rows. Therefore, (4.8) with $[A^N]$ replaced by $[\bar{A}^N]$ again yields (4.9). The rest of the proof remains unchanged. $\square$

Besides for (H4), the VDP-property of the piecewise linear approximations together with the results of the next section will be used to deduce (H5).

4.4. The eigenvalues of the approximate systems. It is known that the spectrum of the generator A coincides with its point spectrum, namely, $\sigma(A) = \{\lambda \in \mathbb{C} \,|\, \det \Delta(\lambda) = 0\}$, where

$$\Delta(\lambda) = \lambda I - A_0 - A_1 e^{-\lambda r}, \qquad \lambda \in \mathbb{C}.$$

The eigenvalues of the finite-dimensional operators A^N are the zeros of $\det(\lambda I^N - [A^N])$ in $\mathbb{C}$. I^N denotes the $n(2N+1)$-identity matrix. To calculate the determinant of the characteristic matrix $\Delta^N(\lambda) = \lambda I^N - [A^N]$ in terms of rational functions of λ, we transform $\Delta^N(\lambda)$, a square matrix of $n \times n$ blocks numbered from zero to $2N$, by elementary row and column operations.

First, to add the odd numbered columns to their succeeding ones, we multiply $\Delta^N(\lambda)$ from the right by S^N, the identity with $(0, I \text{——} 0, I)$ above the diagonal. Then we multiply from the left with T^N, the identity with $-\lambda(\lambda + 6N/r)^{-1}(0, I \text{——} 0, I)$ above the diagonal, in order to subtract $\lambda(\lambda + 6N/r)^{-1}$ times the even numbered rows from their preceding ones. This yields a matrix whose diagonal $4n \times 4n$ blocks look like

$$\begin{bmatrix} q^N(\lambda) & 0 & 0 & 0 \\ 3N/r & \lambda + (6N/r) & 0 & 0 \\ -p^N(\lambda) & 0 & q^N(\lambda) & 0 \\ -3N/r & 0 & 3N/r & \lambda + (6N/r) \end{bmatrix} \otimes I$$

with

$$p^N(\lambda) = \frac{N}{r} - \frac{3N}{r} \lambda \left(\lambda + \frac{6N}{r} \right)^{-1}, \tag{4.12}$$

$$q^N(\lambda) = \lambda + p^N(\lambda).$$

Its zeroth row of blocks is $(\lambda I - A_0, 0\text{——}0, -A_1, 0)$ and its zeroth column reads $(\lambda I - A_0, p^N(\lambda)\, I, (3N/r)I, 0\text{——}0)$. To transform this matrix into a lower triangular matrix let

$$D_N^N = (q^N(\lambda))^{-1} A_1,$$
$$D_j^N = (q^N(\lambda))^{-1} p^N(\lambda) D_{j+1}^N, \qquad j = 1, \cdots, N-1$$

and multiply from the left with D^N, being the identity with $(I, D_1^N, 0, D_2^N, 0\text{——}D_N^N, 0)$ in the first row. Thus

$$\det \Delta^N(\lambda) = (\det q^N(\lambda) I)^N \left(\det\left(\lambda + \frac{6N}{r}\right) I\right)^N \det \Delta_0^N(\lambda) \tag{4.13}$$

where

$$\Delta_0^N(\lambda) = \lambda I - A_0 - p^N(\lambda) D_1^N = \lambda I - A_0 - (r^N(\lambda))^N A_1$$

with

$$r^N(\lambda) = p^N(\lambda)(q^N(\lambda))^{-1}. \tag{4.14}$$

Note that these transformations are possible, as far as $\lambda + 6N/r \neq 0$ and $q^N(\lambda) \neq 0$, i.e., $\lambda \neq -6N/r$ and $\lambda \neq -2N/r \pm i(N/r)\sqrt{2}$, in particular if $\lambda \in \{\lambda \in \mathbb{C} \mid \operatorname{Re} \lambda \geqq -\rho\}$ and $N > \rho r/2$ for some $\rho > 0$. Furthermore we note that if $\operatorname{Re} \lambda \geqq -\rho$, then $\det \Delta^N(\lambda) = 0$ if and only if $\det \Delta_0^N(\lambda) = 0$, provided $N > \rho r/2$.

LEMMA 4.10. (a) *With the projection* $V: M^2 \to \mathbb{R}^n$ *introduced in* (H4) *we have*

$$[V(\lambda I - A^N)^{-1} V^*] = (\Delta_0^N(\lambda))^{-1}, \qquad \lambda \in \rho(A^N).$$

(b) $\Delta_0^N(\lambda) \to \Delta(\lambda)$ *uniformly in* λ *on bounded subsets of* $\mathbb{C}$.

(c) *For any* $\rho > 0$ *let* $K_\rho = \{\lambda \in \mathbb{C} \mid -\rho \leqq \operatorname{Re} \lambda \leqq |A_0| + |A_1|, |\operatorname{Im} \lambda| \leqq |A_0| + 2e^{\rho r}|A_1|\}$. *There exists an* $N(\rho)$ *such that for* $N \geqq N(\rho)$ *all the roots of* $\det \Delta^N(\lambda) = 0$ *with* $-\rho \leqq \operatorname{Re} \lambda$ *lie within* K_ρ.

Proof. Expanding the determinant of $\Delta^N(\lambda)$ by elementary operations, we have seen that $D^N T^N \Delta^N(\lambda) S^N = U^N(\lambda)$, where $U^N(\lambda)$ is a lower triangular block-matrix with $\Delta_0^N(\lambda)$ in the upper left corner. The first column of blocks in $D^N T^N$ and the first row of blocks in $(S^N)^{-1}$ are of the form $(I\ 0\text{——}0)$. Thus, the application of these transformations to the equation $\Delta^N(\lambda)(\psi^0 *\text{——}*)^T = (\phi^0 0\text{——}0)^T$ yields

$$\begin{aligned} U^N(\lambda)(\psi^0 *\text{——}*)^T &= D^N T^N \Delta^N(\lambda) S^N (S^N)^{-1} (\psi^0 *\text{——}*)^T \\ &= D^N T^N (\phi^0 0\text{——}0)^T = (\phi^0 0\text{——}0)^T \end{aligned}$$

or

$$(\psi^0 *\text{——}*)^T = (U^N(\lambda))^{-1} (\phi^0 0\text{——}0)^T \quad \text{for all } \phi^0, \psi^0 \in \mathbb{R}^n.$$

This implies $\psi^0 = (\Delta_0^N(\lambda))^{-1} \phi^0$. But

$$\begin{aligned} V(\lambda I - A^N)^{-1} V^* \phi^0 &= V(\lambda I - A^N)^{-1} (\phi^0, 0) = [V](\Delta^N(\lambda))^{-1} (\phi^0 0\text{——}0)^T \\ &= [V](\psi^0 *\text{——}*)^T = \psi^0 = (\Delta_0^N(\lambda))^{-1} \phi^0 \end{aligned}$$

for all $\phi^0 \in \mathbb{R}^n$, and this proves (a).

From (4.12), (4.14) we have

$$r^N(\lambda) = \frac{6N^2 - 2Nr\lambda}{6N^2 + 4Nr\lambda + \lambda^2 r^2} = \left(1 + \frac{r\lambda}{N}\right)^{-1} \left(1 + \frac{3r^2\lambda^2}{6N^2 + 4Nr\lambda - 2r^2\lambda^2}\right)^{-1}.$$

As $N \to \infty$

$$\left(1 + \frac{r\lambda}{N}\right)^{-N} \to e^{-r\lambda} \quad \text{and} \quad \left(1 + \frac{3r^2\lambda^2}{6N^2 + 4Nr\lambda - 2r^2\lambda^2}\right)^{-N} \to 1$$

uniformly in λ on bounded subsets of $\mathbb{C}$, and (b) follows. To prove (c) let $\operatorname{Re}\lambda \geqq -\rho$ and $\det \Delta^N(\lambda) = 0$. For $N > \rho r/2$ it follows that $\det \Delta_0^N(\lambda) = 0$. Thus λ is an eigenvalue of $A_0 + (r^N(\lambda))^N A_1$, which implies $|\lambda| \leqq |A_0| + |(r^N(\lambda))^N||A_1|$. For $\operatorname{Re}\lambda \geqq 0$, $|r^N(\lambda)| \leqq 1$. Computing $|r^N(\lambda)|^2$ and its derivative with respect to $\operatorname{Re}\lambda$ shows $|r^N(\lambda)| \leqq |r^N(-\rho)|$ for $\operatorname{Re}\lambda \in [-\rho, 0]$ and N sufficiently large. Hence the bounds in K_ρ are obtained from the convergence $(r^N(-\rho))^N \to e^{\rho r}$. □

Although the estimates in (c) are not tight, they lead to precise results by use of the following consequence of Rouché's Theorem.

PROPOSITION 4.11. *Let f, f^N, $N = 1, 2, \cdots$ be holomorphic inside and on a closed bounded contour $\Gamma \subseteq \mathbb{C}$. If f has no zeros on Γ and if $f^N \to f$ uniformly on Γ, then there exists an $N_0 \in \mathbb{N}$ such that for $N \geqq N_0$, f^N and f have the same number of zeros (counted according to their multiplicities) inside Γ.*

LEMMA 4.12. (a) *If λ_0 is an eigenvalue of A with multiplicity k, then for any $\varepsilon > 0$ (small enough) there is an N_0 such that each A^N, $N \geqq N_0$, possesses k eigenvalues in $B(\lambda_0, \varepsilon) = \{\lambda \in \mathbb{C} \,|\, |\lambda - \lambda_0| < \varepsilon\}$.*

(b) *Let $\rho > 0$, $G_\rho = \{\lambda \in \mathbb{C} \,|\, \operatorname{Re}\lambda \geqq -\rho\}$ and λ_i, $i = 1, \cdots, l$ be the eigenvalues of A in G_ρ. For any $\varepsilon > 0$ (small enough) there exists N_0 such that the operators A^N, $N \geqq N_0$, have no eigenvalues in $G_\rho \setminus \bigcup_{i=1}^l B(\lambda_i, \varepsilon)$.*

Proof. Assume that λ_0 is the only zero of $\det \Delta(\lambda)$ in $\overline{B(\lambda_0, \varepsilon)}$. $\partial B(\lambda_0, \varepsilon)$ is bounded and $\det \Delta_0^N(\lambda) \to \det \Delta(\lambda)$ uniformly on bounded sets. Thus, (a) follows at once from Proposition 4.11.

Choose, without loss of generality, $\varepsilon > 0$ such that $\det \Delta(\lambda)$ has no zero in $\partial G_\rho \setminus \bigcup_{i=1}^l \partial B(\lambda_i, \varepsilon)$. We know from (4.13) that, if $N > \rho r/2$, $\lambda \in G_\rho$ is an eigenvalue of A^N if and only if $\det \Delta_0^N(\lambda) = 0$. Write $G_\rho \setminus \bigcup_{i=1}^l B(\lambda_i, \varepsilon) = G_1 \cup G_2$, where

$$G_1 = (G_\rho \cap K_\rho) \setminus \bigcup_{i=1}^{l} B(\lambda_i, \varepsilon), \qquad G_2 = (G_\rho \cap K_\rho^c) \setminus \bigcup_{i=1}^{l} B(\lambda_i, \varepsilon).$$

G_1 is bounded and ∂G_1 contains no zero of $\det \Delta(\lambda)$. Thus, there is an $N_1 > \rho r/2$ such that for all $N \geqq N_1$ $\det \Delta_0^N(\lambda)$ has as many zeros in G_1 as $\det \Delta(\lambda)$, that is, $\det \Delta_0^N(\lambda)$ has no zero in G_1. Since $G_2 \subseteq K_\rho^c$, there is no eigenvalue of A^N in G_2, if N is sufficiently large. □

This shows that the eigenvalues of A are approximated by the eigenvalues of the operators A^N. Moreover, given $\varepsilon > 0$, let $\rho \in \mathbb{R}$ and λ_i, $i = 1, \cdots, l$, be the eigenvalues of the hereditary system with $\operatorname{Re}\lambda_i \geqq \rho$. Then the piecewise linear approximations do not have eigenvalues in the right halfplane $\operatorname{Re}\lambda \geqq \rho$ outside the balls $B(\lambda_i, \varepsilon)$, $i = 1, \cdots, l$, provided N is sufficiently large.

4.5. Uniform stability. From the results of the previous section, we conclude that if $S(\cdot)$ is stable, i.e., $\|S(t)\| \leqq M e^{-\omega_0 t}$ with some M, $\omega_0 > 0$, then for all $\omega < \omega_0$ there are an N_ω and constants M_N such that for all $N \geqq N_\omega$

$$\|S^N(t)\| \leqq M_N e^{-\omega t}. \tag{4.15}$$

To get uniformity with respect to N on the right-hand side of these estimates, we follow an idea given by Ito [11] in connection with his Legendre-tau approximations. The idea is to establish uniformity in (4.15) for one special no-delay case and to interpret this special case as a perturbation of the general situation.

Let us consider the equation $\dot{x}(t) = -x(t)$, $t \geqq 0$, in $\mathbb{R}^n$ as if it were a functional differential equation with delay, demanding the initial condition $(x(0), x_0) = \phi \in M^2$. We approximate by our piecewise linear scheme, denoting the approximating generators by A_0^N. They are given in Definition 4.3 with $A_0 = -I$, $A_1 = 0$. The representation $[A_0^N]$

is a lower triangular block matrix with $-I$ in the left upper position. Therefore the first row of blocks in $e^{[A_0^N]t}$ is given by $(e^{-t}I\ 0\text{——}0)$. Hence, for all $\phi \in Z^N$

$$|Ve^{A_0^N t}\phi| = |\alpha_0(e^{[A_0^N]t}\alpha^N(\phi))| = |e^{-t}\alpha_0(\phi)| = e^{-t}|\phi^0| \leqq e^{-t}\|\phi\|.$$

Thus, by Lemma 4.8, there is a $c>0$ such that

$$\int_0^\infty \|e^{A_0^N t}\phi\|^2\, dt \leqq c\|\phi\|^2, \qquad \phi \in Z^N \tag{4.16}$$

and by Lemma 3.2 there exist constants M_0, $\alpha_0 > 0$ such that for all $N = 1, 2, \cdots$

$$\|e^{A_0^N t}|_{Z^N}\| \leqq M_0\, e^{-\alpha_0 t}, \qquad t \geqq 0.$$

Remark. Guided by these arguments, we easily see that the spline approximation scheme presented in [13], [14] does not have the VDP-property. Because the first row of blocks of the matrix representing the spline generators ${}^sA_0^N$ is also of type $(-I\ 0\text{——}0)$ we obtain for all ϕ in the spline subspace ${}^sZ^N$, $|(e^{{}^sA_0^N t}\phi)^0| \leqq e^{-t}\|\phi\|$, as above. Thus, if a spline analogue of Lemma 4.8 holds, then by Lemma 3.2 $\|e^{{}^sA_0^N t}|_{{}^sZ^N}\| \leqq K e^{-\varepsilon t}$, for some K, $\varepsilon > 0$, independent of N. But this contradicts the peculiar eigenvalue behavior of the spline scheme (see [14, Prop. 4.6]).

LEMMA 4.13. *If $\|S(t)\| \leqq M e^{-\omega_0 t}$, $t \geqq 0$ for some M, $\omega_0 > 0$ then for all $\omega < \omega_0$ there exist N_ω and $\tilde{M}$ such that for all $N \geqq N_\omega$*

$$\|S^N(t)|_{Z^N}\| \leqq \tilde{M} e^{-\omega t}, \qquad t \geqq 0.$$

Proof. Let $0 < \omega < \omega_0$. We have seen in the proof of Lemma 4.10 that there is a constant c such that $|A_0 + (r^N(-\omega + i\tau))^N A_1| \leqq c$, $\tau \in \mathbb{R}$, if $N \geqq N(\omega)$. It follows that

$$|(\Delta_0^N(-\omega + i\tau))^{-1}| \leqq \frac{1}{|-\omega + i\tau| - c} \quad \text{for } |-\omega + i\tau| > c$$

(det $\Delta_0^N(-\omega + i\tau) \neq 0$ if N is large enough). Thus, from the uniform convergence of $\Delta_0^N(\lambda)$ to $\Delta(\lambda)$ on the set $\{\lambda = -\omega + i\tau \,|\, |\lambda| \leqq c\}$, we have

$$|(\Delta_0^N(-\omega + i\tau))^{-1}| \leqq \gamma, \qquad \tau \in \mathbb{R}, \quad N \geqq N_\omega \tag{4.17}$$

with some $N_\omega \in \mathbb{N}$ and $\gamma > 0$.

Define $A_\omega^N = \omega I + A^N$ and let $\phi \in Z^N$. The trajectory $e^{A_0^N t}\phi$ is the solution of

$$\dot{z}(t) = A_0^N z(t) = A_\omega^N z(t) + (A_0^N - A_\omega^N)z(t), \qquad t \geqq 0, \quad z(0) = \phi.$$

Equivalently,

$$e^{A_0^N t}\phi = e^{A_\omega^N t}\phi + \int_0^t e^{A_\omega^N (t-s)}(A_0^N - A_\omega^N)\, e^{A_0^N s}\phi\, ds;$$

hence

$$e^{A_\omega^N t}\phi = e^{A_0^N t}\phi + \int_0^t e^{A_\omega^N (t-s)} V^* f^N(s, \phi)\, ds$$

where $f^N(\cdot, \phi): \mathbb{R} \to \mathbb{R}^n$ is given by $f^N(s, \phi) = F e^{A_0^N s}\phi$, $s \geqq 0$ with $F: M^2 \to \mathbb{R}^n$:

$$F\psi = (I + \omega I + A_0)\psi^0 + A_1\psi^1(-r) \tag{4.18}$$

because the L^2-component of $(A_\omega^N - A_0^N)\psi$ vanishes for all $\psi \in Z^N$. From (4.18) and (4.16) it is clear that

$$\|f^N(\cdot, \phi)\|_{L^2(0,\infty;\mathbb{R}^n)} \leqq \kappa \|\phi\|_{M^2} \tag{4.19}$$

for some $\kappa > 0$ not depending on N. We write

$$Ve^{A_\omega^N t}\phi = Ve^{A_0^N t}\phi + y(t), \qquad t \geqq 0 \tag{4.20}$$

with

$$y(t) = \int_0^t Ve^{A_\omega^N (t-s)} V^* f^N(s, \phi)\, ds, \qquad t \geqq 0.$$

Letting $f^N(s, \phi) = 0$, $s \leqq 0$ and $e^{A_\omega^N (t-s)} = 0$, $s > t$ or $s \leqq 0$, we have $f^N(\cdot, \phi) \in L^1(\mathbb{R}; \mathbb{R}^n) \cap L^2(\mathbb{R}; \mathbb{R}^n)$ and by (4.15) $e^{A_\omega^N \cdot} \in L^1(\mathbb{R}; \mathscr{L}(Z^N) \cap L^2(\mathbb{R}; \mathscr{L}(Z^N))$, if N is sufficiently large. The calculation of the Fourier-transform $\hat{y}(\cdot)$ of the convolution $y(\cdot)$ yields

$$\begin{aligned}
\hat{y}(\tau) &= (Ve^{A_\omega^N \cdot} V^*)\hat{}(\tau)(f^N(\cdot, \phi))\hat{}(\tau) \\
&= \int_{-\infty}^{\infty} e^{-i\tau t} Ve^{A_\omega^N t} V^*\, dt (f^N(\cdot, \phi))\hat{}(\tau) \\
&= V(-\omega + i\tau - A^N)^{-1} V^* (f^N(\cdot, \phi))\hat{}(\tau), \qquad \tau \in \mathbb{R}.
\end{aligned}$$

By Plancherel's Theorem and (4.19), we get

$$\int_{-\infty}^{\infty} |(f^N(\cdot, \phi))\hat{}(\tau)|^2\, d\tau = \int_0^{\infty} |f^N(t, \phi)|^2\, dt \leqq \kappa^2 \|\phi\|^2.$$

Thus

$$\begin{aligned}
\int_0^{\infty} |y(t)|^2\, dt &= \int_{-\infty}^{\infty} |\hat{y}(\tau)|^2\, d\tau \\
&\leqq \int_{-\infty}^{\infty} |V(-\omega + i\tau - A^N)^{-1} V^*|^2 |(f^N(\cdot, \phi))\hat{}(\tau)|^2\, d\tau \\
&= \int_{-\infty}^{\infty} |(\Delta_0^N(-\omega + i\tau))^{-1}|^2 |(f^N(\cdot, \phi))\hat{}(\tau)|^2\, d\tau
\end{aligned}$$

by Lemma 4.10(a). Therefore, from (4.17)

$$\int_0^{\infty} |y(t)|^2\, dt \leqq \gamma^2 \kappa^2 \|\phi\|^2, \qquad N \geqq N_\omega$$

and, by (4.20), (4.16), and Lemma 4.8,

$$\int_0^{\infty} \|e^{A_\omega^N t}\phi\|^2\, dt \leqq \text{const.}\, \|\phi\|^2, \qquad N \geqq N_\omega, \quad \phi \in Z^N.$$

Lemma 3.2 thus yields constants $\tilde{M}$, $\varepsilon > 0$ not depending on N, such that

$$e^{\omega t} \|e^{A^N t}|_{Z^N}\| = \|e^{A_\omega^N t}|_{Z^N}\| \leqq \tilde{M} e^{-\varepsilon t}, \qquad N \geqq N_\omega. \qquad \square$$

So, if the hereditary semigroup $S(\cdot)$ is exponentially stable with some decay rate ω_0, it is approximated by piecewise linear systems with decay rate arbitrarily close to ω_0.

Another immediate consequence of Lemma 4.13 is the following corollary.

COROLLARY 4.14. *The semigroups* $\tilde{S}^N(\cdot)$ *generated by* $\tilde{A}^N p^N = (A^N - BR^{-1}B^*\Pi)p^N$ *satisfy hypothesis* (H5).

Proof. If the hereditary system is stabilizable, there exists a solution Π of the algebraic Riccati equation, and the closed-loop semigroup $\bar{S}(\cdot)$ generated by $\bar{A} = A - BR^{-1}B^*\Pi$ is exponentially stable. The approximations to $\bar{S}(\cdot)$ are actually given by $\tilde{S}^N(\cdot)$. Thus the previous lemma yields constants $\tilde{M}$, $\omega > 0$ such that

$$\|\tilde{S}^N(t)|_{Z^N}\| \leqq \tilde{M} e^{-\omega t}, \qquad t \geqq 0$$

for all N sufficiently large. □

Summarizing, we have proved that the convergence statements of Theorem 3.3 are valid when the piecewise linear approximation scheme is applied to infinite time horizon hereditary control problems.

4.6. Multiple and distributed delay terms. The piecewise linear approximation scheme can also be used for systems with more than one discrete delay ($p \geqq 2$ in (1.1)) and/or distributed delays ($A_{01}(\cdot) \not\equiv 0$, $A_{01} \in L^2(-h, 0; \mathbb{R}^{n\times n})$ in (1.1)). Then the intervals $I_k = [-h_k, -h_{k-1})$, $I = [-h_1, 0]$ of length $r_k = h_k - h_{k-1}$ are each divided into N subintervals of length r_k/N and the $(2npN)$-dimensional space Y^N consists of functions that are polynomials of degree one on each of these subintervals. In fact, all the results stated above for the single-delay case are valid in the general case. The proofs can be found in [17], which is a previous version of the present paper.

The arguments are essentially the same as above, with the main modifications arising from two facts: in the general case dom A^* contains pairs (ψ^0, ψ^1) with discontinuous functions ψ^1, the jumps at $-h_k$ given by $A_k^T\psi^0$; in order to get dissipativity of A and A^N, an equivalent inner product placing increasing weights on the intervals I_{p-k} is employed (cf. [13, Lemma 5.4]).

The generators A^N have δ^{N-} terms at all the discontinuities in the subspaces Y^N, while the adjoints A^{N*} in addition must reflect the jumps of elements in dom A^* at the points $-h_k$ and so the jump heights $A_k^T\psi^0$ appear in the arguments of the δ^{N+} terms for these points. $A^{N*}\psi \to A^*\psi$ for ψ in an appropriate set $D^* \subseteq$ dom A^* follows from $(A_{01}^T\psi^0)^N \to A_{01}^T\psi^0$ for $A_{01} \in L^2(-h, 0; \mathbb{R}^{n\times n})$; the density of $(\lambda - A^*)D^*$ can be shown by arguments similar to those in the proof of Theorem 7.2 of [9]. This avoids additional smoothness assumptions on A_{01} as it is required in [11]–[13].

The coupling of the intervals I_k as reflected in the matrices $[A^N]$ again yields the VDP-property. Expanding $\det \Delta^N(\lambda)$ for the general case and the determination of bounds on its zeros is laborious but leads to the above stated conclusions on the eigenvalues and the uniform stability of the approximate systems.

For the numerical implementation of a problem with multiple and/or distributed delays we must use the $n(2pN+1)$ matrices:

$$Q^N = \operatorname{diag}\left(1, \frac{r_1}{N}, \frac{r_1}{3N}, \cdots, \frac{r_1}{N}, \frac{r_1}{3N}, \frac{r_2}{N}, \frac{r_2}{3N}, \cdots, \frac{r_p}{N}, \frac{r_p}{3N}\right) \otimes I$$

and H^N, obtained by replacing the first row of $n \times n$ blocks in (4.7) by the row of blocks $(A_0, A_1^N \text{——} A_p^N)$ with

$$A_k^N = (A_{k1}^N, A_{k2}^N \cdots A_{k,2N-2}^N, A_{k,2N-1}^N + A_k, A_{k,2N}^N - A_k),$$

$$A_{kj}^N = \int_{-h}^{0} A_{01}(s)\, e_{kj}^N(s)\, ds, \qquad k = 1, \cdots, p, \quad j = 1, \cdots, 2N,$$

where e_{kj}^N are the basis elements of Y^N, constructed for each interval I_k by analogy to the single-delay case.

4.7. Examples. Testing the numerical performance of the approximation scheme that has been developed in this paper, we have employed it in several examples and

compared the outcomes with the results produced by other schemes. As far as these examples are representative, it turns out that the piecewise linear and the first-order spline approximations [13] are of the same numerical accuracy (for the same approximation index N), but both are inferior to the Legendre methods [10], [12].

In the case of the finite time horizon the Riccati differential equation

$$\frac{d}{dt}[\Pi^N(t)]+[(A^N)^*][\Pi^N(t)]+[\Pi^N(t)][A^N]$$
$$-[\Pi^N(t)][B^N]R^{-1}[B^N]^T[\Pi^N(t)]+[W^N]=0, \qquad 0\leqq t\leqq T,$$
$$[\Pi^N(T)]=[G^N]$$

is transformed, by taking $\Gamma^N(t)=Q^N[\Pi^N(T-t)]$, into a standard Riccati matrix differential equation

$$\frac{d}{dt}\Gamma^N(t)+[A^N]^T\Gamma^N+\Gamma^N[A^N]-\Gamma^N[B^N]R^{-1}[B^N]^T\Gamma^N+[W^N]=0,$$

(4.21) $$0\leqq t\leqq T,$$

$$\Gamma^N(0)=[G^N].$$

Observe that the self-adjointness of $\Pi^N(t)$ implies $[\Pi^N(t)]^TQ^N=Q^N[\Pi^N(t)]$ and hence $\Gamma^N(t)^T=\Gamma^N(t)$. Since B^N, W^N, G^N refer exclusively to the $\mathbb{R}^n$-component of Z^N, we have

$$[B^N]=\operatorname{col}(B_0,0,\cdots,0)\in\mathbb{R}^{n(2pN+1)\times m},$$

$$[W^N]=\begin{bmatrix} W_0 & 0 & \cdots & 0\\ 0 & & & 0\\ \vdots & & & \vdots\\ 0 & & & 0\end{bmatrix}\in\mathbb{R}^{n(2pN+1)\times n(2pN+1)},$$

$$[G^N]=\begin{bmatrix} G_0 & 0 & \cdots & 0\\ 0 & & & 0\\ \vdots & & & \vdots\\ 0 & & & 0\end{bmatrix}\in\mathbb{R}^{n(2pN+1)\times n(2pN+1)}.$$

Thus (if $p=1$, $A_{01}\equiv 0$) we can reduce the dimension of (4.21) introducing the $2n\times n(2N+1)$ matrices

$$F_1^N=\begin{bmatrix} F_0 & 0\cdots 0 & G_0A_1-G_0A_1\\ A_1^TG_0 & 0\cdots\cdots 0\end{bmatrix},\qquad F_2^N=\begin{bmatrix} I & 0\cdots\cdots 0\\ 0\cdots\cdots 0 & I-I\end{bmatrix}$$

where F_0 is given by

$$F_0=A_0^TG_0+G_0A_0-G_0B_0R^{-1}B_0^TG_0+W_0,$$

in order to get the factorization

$$\dot\Gamma(0)=(F_1^N)^T(F_2^N).$$

This implies ([18, p. 304 ff.])

(4.22) $$\Gamma^N(t)=[G^N]+\int_0^t L_1^N(s)^TL_2^N(s)\,ds,$$

$L_i^N(t)$ being the solution of

$$\frac{d}{dt}L_i^N(t) = L_i^N(t)([A^N]-[B^N]R^{-1}[B^N]^T\Gamma^N(t)), \tag{4.23}$$

$$L_i^N(0) = F_i^N, \qquad i=1,2.$$

Multiplying (4.22) from the left with $[B^N]^T$, we obtain

$$[B^N]^T\dot{\Gamma}^N(t) = [B^N]^TL_1^N(t)^TL_2^N(t), \qquad 0 \leqq t \leqq T, \tag{4.24}$$

$$[B^N]^T\Gamma^N(0) = [B^N]^T[G^N].$$

Solving the $n(4n+m)(2N+1)$ differential equations (4.23), (4.24), we get $[B^N]^T\Gamma^N(t)$. But this is all we need for the computation of the suboptimal control $\hat{u}^N(t)$ (see (3.1)). Denoting the $m \times n$ blocks in $[B^N]^T[\Pi^N(t)]$ by $\beta_0^N(t), \cdots, \beta_{2N}^N(t)$, we have

$$\hat{u}^N(t) = -R^{-1}\left\{\beta_0^N(t)\hat{x}^N(t) + \sum_{j=1}^{2N}\int_{-r}^{0}\beta_j^N(t)\, e_{1j}^N(s)\hat{x}^N(t+s)\, ds\right\}, \tag{4.25}$$

$$0 \leqq t \leqq T$$

where $\hat{x}^N(t)$ is the solution of

$$\dot{x}(t) = A_0x(t) + A_1x(t-r) + B_0\hat{u}^N(t) \tag{4.26}$$

in $\mathbb{R}^n$. In each term of the sum in (4.25), the integration ranges only over one of the intervals I_j^N, $j=1, \cdots, N$.

Numerically the systems (4.25), (4.26) were solved simultaneously by an appropriately adjusted fourth-order Runge-Kutta procedure combined with Simpson's rule for the evaluation of the integrals.

Example 4.1. Minimize

$$J(u) = \frac{1}{2}(x_1(2)^2 + x_2(2)^2) + \frac{1}{2}\int_0^2 (u_1(t)^2 + u_2(t)^2)\, dt$$

subject to

$$\dot{x}(t) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} x(t-1) + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} u(t), \qquad 0 \leqq t \leqq 2,$$

$$x(0) = x_0(t) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad -1 \leqq t \leqq 0.$$

The true solutions $\bar{u}(t)$, $\bar{x}(t)$ (see [2]) and the piecewise linear approximations with index $N=4$, 8, 16 are presented in Tables 4.1 and 4.2. The relatively greatest errors occur around $t=1$ and $t=2$, where the derivatives of $\bar{x}(t)$ and $\bar{u}(t)$ have jumps, while $\hat{x}^N(t)$ and $\hat{u}^N(t)$ are of course continuously differentiable.

For the infinite time horizon control problem ($p=1$, $A_{01} \equiv 0$) the suboptimal control and trajectory are again calculated via (4.25), (4.26) when $\Pi^N(t)$ is replaced by the stationary operator Π^N, that is, the solution of the algebraic Riccati equation (3.3). The transformation $\Gamma^N = Q^N[\Pi^N]$ yields a standard Riccati matrix equation

$$[A^N]^T\Gamma^N + \Gamma^N[A^N] - \Gamma^N[B^N]R^{-1}[B^N]^T\Gamma^N + [W^N] = 0$$

in $\mathbb{R}^{n(2N+1)\times n(2N+1)}$, which was solved by the Newton-Kleinman Algorithm as presented in [18]. In each step of this algorithm, a Lyapunov matrix equation was solved using

TABLE 4.1

t \ $\hat{u}_1$	$\hat{u}_1^4(t)$	$\hat{u}_1^8(t)$	$\hat{u}_1^{16}(t)$	$\bar{u}_1(t)$
0	−1.0602	−1.0599	−1.0598	−1.0598
0.25	−0.8419	−0.8419	−0.8419	−0.8419
0.5	−0.6209	−0.6241	−0.6239	−0.6239
0.75	−0.4008	−0.4030	−0.4060	−0.4060
1.0	−0.2268	−0.2116	−0.2022	−0.1880
1.25	−0.1743	−0.1860	−0.1884	|
1.5	−0.1897	−0.1878	−0.1880	|
1.75	−0.1877	−0.1880	−0.1880	|
2.0	−0.1880	−0.1880	−0.1880	−0.1880
t \ $\hat{u}_2$	$\hat{u}_2^4(t)$	$\hat{u}_2^8(t)$	$\hat{u}_2^{16}(t)$	$\bar{u}_2(t)$
0	−0.8721	−0.8719	−0.8718	−0.8718
0.25	−0.8721	|	|	|
0.5	−0.8721	|	|	|
0.75	−0.8721	−0.8719	|	|
1.0	−0.8720	−0.8718	|	|
1.25	−0.8720	|	|	|
1.5	−0.8719	|	|	|
1.75	−0.8718	|	|	|
2.0	−0.8719	−0.8718	−0.8718	−0.8718
$J(\hat{u})$	1.4018	1.4017	1.4017	1.4017

TABLE 4.2

t \ $\hat{x}_1$	$\hat{x}_1^4(t)$	$\hat{x}_1^8(t)$	$\hat{x}_1^{16}(t)$	$\bar{x}_1(t)$
0.25	0.76221	0.76227	0.76228	0.76229
0.5	0.57927	0.57902	0.57905	0.57906
0.75	0.45179	0.45050	0.45031	0.45032
1.0	0.37519	0.37582	0.37500	0.37607
1.25	0.32780	0.32910	0.32905	0.32906
1.5	0.28219	0.28203	0.28205	0.28208
1.75	0.23498	0.23503	0.23504	0.23504
2.0	0.18799	0.18802	0.18803	0.18803
t \ $\hat{x}_2$	$\hat{x}_2^4(t)$	$\hat{x}_2^8(t)$	$\hat{x}_2^{16}(t)$	$\bar{x}_2(t)$
0.25	1.03198	1.03203	1.03205	1.03205
0.5	1.06395	1.06407	1.06409	1.06410
0.75	1.09593	1.09610	1.09614	1.09615
1.0	1.12791	1.12814	1.12818	1.12821
1.25	1.12906	1.12933	1.12934	1.12941
1.5	1.07761	1.07790	1.07797	1.07799
1.75	0.98737	0.98748	0.98755	0.98758
2.0	0.87187	0.87180	0.87179	0.87180

the quadratic procedure given by Smith (see also [18, p. 297]). The time-independent $m \times n$ blocks β_j^N, $j = 0, \cdots, 2N$ were then employed in (4.25). Furthermore, with

$$[\Pi^N] = \begin{bmatrix} \Pi_{00}^N \Pi_{11}^N \cdots \Pi_{1,2N}^N \\ * \text{———} * \\ | \qquad\qquad | \\ * \text{———} * \end{bmatrix}$$

we give some values of the feedback kernel

$$\Pi_1^N(s) = \sum_{j=1}^{2N} \Pi_{1j}^N e_j^N(s),$$

which together with Π_{00}^N determines the feedback law of the Nth approximation. At the meshpoints, we simply have

$$\Pi_1^N(0) = \Pi_{11}^N + \Pi_{12}^N, \quad \Pi_1^N(t_j^N) = \Pi_{1,2j-1}^N - \Pi_{1,2j}^N, \quad j = 1, \cdots, N.$$

Example 4.2. This is the problem of minimizing

$$J(u) = \int_0^\infty [x(t)^2 + u(t)^2]\, dt$$

subject to

$$\dot{x}(t) = x(t) + x(t-1) + u(t), \qquad t \geqq 0,$$
$$x(0) = 0, \quad x_0(t) = \sin \pi t, \quad -1 \leqq t \leqq 0.$$

Table 4.3 gives the optimal costs $J^N = \langle (x(0), x_0), \Pi^N(x(0), x_0) \rangle_{M^2}$ of the approximating systems and the costs $J(\hat{u}^N)$ when the original system is controlled by $\hat{u}^N$. Table 4.4 presents the feedback gains Π_{00}^N and $\Pi_1^N(s)$ at the meshpoints $-j/4$, $j = 0, \cdots, 4$. In Table 4.5, we list the values of $\hat{u}^N(j/4)$, $\hat{x}^N(j/4)$, $j = 0, \cdots, 12$.

TABLE 4.3

N	J^N	$J(\hat{u}^N)$
4	0.32117	0.32143
8	0.32138	0.32143
16	0.32142	0.32143

TABLE 4.4

	Π_{00}^4	Π_{00}^8	Π_{00}^{16}
	2.8083	2.8092	2.8094
j	$\Pi_1^4(-j/4)$	$\Pi_1^8(-j/4)$	$\Pi_1^{16}(-j/4)$
0	0.6349	0.6365	0.6369
1	0.8700	0.8801	0.8838
2	1.2544	1.2745	1.2803
3	1.8518	1.8812	1.8895
4	2.7507	2.7930	2.8050

TABLE 4.5

t \ $\hat{u}$	$\hat{u}^4(t)$	$\hat{u}^8(t)$	$\hat{u}^{16}(t)$
0	0.86968	0.86836	0.86818
0.25	0.64873	0.64888	0.64891
0.5	0.49572	0.49647	0.49657
0.75	0.36350	0.36395	0.36400
1.0	0.24664	0.24624	0.24618
1.25	0.16200	0.16153	0.16147
1.5	0.11027	0.10997	0.10993
1.75	0.08030	0.08022	0.08021
2.0	0.06013	0.06015	0.06015
2.25	0.04349	0.04347	0.04347
2.5	0.02988	0.02982	0.02982
2.75	0.02002	0.01996	0.01995
3.0	0.01377	0.01373	0.01372
t \ $\hat{x}$	$\hat{x}^4(t)$	$\hat{x}^8(t)$	$\hat{x}^{16}(t)$
0.25	0.11276	0.11260	0.11258
0.5	0.05337	0.05332	0.05332
0.75	−0.06642	−0.06628	−0.06626
1.0	−0.10870	−0.10849	−0.10846
1.25	−0.06170	−0.06160	−0.06158
1.5	−0.01396	−0.01396	−0.01397
1.75	0.00758	0.00753	0.00752
2.0	0.00180	0.00178	0.00178
2.25	−0.00787	−0.00784	−0.00784
2.5	−0.01034	−0.01030	−0.01029
2.75	−0.00648	−0.00646	−0.00646
3.0	−0.00178	−0.00178	−0.00178

Example 4.3. A simplified model for a wind tunnel at the NASA Langley Research Center is given by (see [4])

$$\dot{x}(t)=\begin{bmatrix}-a & 0 & 0\\ 0 & 0 & 1\\ 0 & -\omega^2 & -2\xi\omega\end{bmatrix}x(t)+\begin{bmatrix}0 & ka & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}x(t-0.33)+\begin{bmatrix}0\\ 0\\ -\omega^2\end{bmatrix}u(t), \qquad t\geqq 0, \tag{4.27}$$

$x(0)=\operatorname{col}(-0.1, 8.547, 0)\equiv x_0(t)$, $-0.33\leqq t\leqq 0$, where $k=-0.0117$, $\xi=0.8$, $\omega=6.0$, $1/a=1.964$. We want to minimize

$$J(u)=\int_0^\infty [10^4 x_1(t)^2+u(t)^2]\,dt$$

subject to (4.27).

The true solution of the problem has been given in [15]. Note that the matrix W_0 weighting the contribution of the state trajectory to the costs is singular in this example,

in contrast to the assumptions in Theorem 3.3. However, the piecewise linear approximation scheme produces the following values for $J(\hat{u}^N)$ and J^N (Table 4.6).

In Table 4.7, we compare the first block of the Riccati matrix Π^N with the $\mathbb{R}^3$-component of Π. The matrices $\Pi_1(t)$ and $\Pi_1^N(t)$, $-r=-0.33\leqq t\leqq 0$, have nonzero entries only in their second columns, which are shown in Table 4.8 for $t=-jr/4$, $j=0,\cdots,4$.

TABLE 4.6

N	$J(\hat{u}^N)$	J^N
4	136.4490	136.1785
8	136.4490	136.2921
16	136.4493	136.3486
$J(\bar{u})$	136.4049	

TABLE 4.7

N	Π_{00}^N		
4	8677.02161	−9.81498	−0.94768
	−9.81498	0.01850	0.00186
	−0.94768	0.00186	0.00019
8	8677.02502	−9.81503	−0.94768
	−9.81503	0.01851	0.00186
	−0.94768	0.00186	0.00019
16	8677.02551	−9.81504	−0.94768
	−9.81504	0.01851	0.00186
	−0.94768	0.00186	0.00019
Π_{00}	8677.02405	−9.81505	−0.94768
	−9.81505	0.01851	0.00186
	−0.94768	0.00186	0.00019

TABLE 4.8

j	0	1	2	3	4
$\Pi_1^4(-jr/4)$	−41.39647	−43.83755	−46.36726	−48.97855	−51.67634
	0.06915	0.06653	0.06358	0.06095	0.05845
	0.00669	0.00641	0.00614	0.00589	0.00564
$\Pi_1^8(-jr/4)$	−41.39710	−43.84694	−46.37700	−48.98892	−51.68730
	0.06917	0.06633	0.06359	0.06097	0.05847
	0.00668	0.00640	0.00614	0.00589	0.00565
$\Pi_1^{16}(-jr/4)$	−41.39721	−43.84929	−46.37952	−48.99157	−51.69010
	0.06917	0.06631	0.06360	0.06098	0.05847
	0.00668	0.00640	0.00614	0.00589	0.00565
$\Pi_1(-jr/4)$	−41.39721	−43.85008	−46.38034	−48.99246	−51.69103
	0.06917	0.06632	0.06360	0.06098	0.05847
	0.00668	0.00641	0.00614	0.00589	0.00565

Acknowledgment. I thank Professor F. Kappel for numerous discussions, valuable hints, and the extensive care concerning my work on this paper.

REFERENCES

[1] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximation*, SIAM J. Control Optim., 16 (1978), pp. 169-208.

[2] H. T. BANKS, J. A. BURNS, E. M. CLIFF, AND P. R. THRIFT, *Numerical solutions of hereditary control problems via an approximation technique*, LCDS Technical Report 15-6, Brown University, Providence, RI, 1975.

[3] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496-522.

[4] H. T. BANKS, G. I. ROSEN, AND K. ITO, *A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 830-855.

[5] J. A. BURNS AND E. M. CLIFF, *Methods for approximating solutions to linear hereditary quadratic optimal control problems*, IEEE Trans. Automat. Control, 23 (1978), pp. 21-36.

[6] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, Berlin, 1978.

[7] R. DATKO, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428-445.

[8] J. S. GIBSON, *The Riccati integral equations for optimal control problems in Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537-565.

[9] ———, *Linear-quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95-139.

[10] K. ITO AND R. TEGLAS, *Legendre-tau approximation for functional differential equations, Part* II: *The linear quadratic optimal control problem*, SIAM J. Control Optim., 25 (1987), pp. 1379-1408.

[11] K. ITO, *Legendre-tau approximation for functional differential equations, Part* III: *Eigenvalue approximations and uniform stability*, in Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Springer-Verlag, Heidelberg, 1985, pp. 191-212.

[12] F. KAPPEL AND G. PROPST, *Approximation of feedback controls for delay systems using Legendre polynomials*, Confer. Sem. Mat. Univ. Bari, 201 (1984), pp. 1-36.

[13] F. KAPPEL AND D. SALAMON, *Spline approximations for retarded systems and the Riccati equation*, SIAM J. Control Optim., 25 (1987), pp. 1082-1117.

[14] ———, *On the stability properties of spline approximations for retarded systems*, SIAM J. Control Optim., 27 (1989), pp. 407-431.

[15] A. MANITIUS AND H. TRAN, *Numerical simulation of a nonlinear feedback controller for a wind tunnel model involving a time delay*, Optimal Control Appl. Methods, 7 (1986), pp. 19-39.

[16] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[17] G. PROPST, *Piecewise linear approximation for hereditary control problems*, ICASE Report 87-11, NASA Langley Research Center, Hampton, VA, NASA CR-178260.

[18] D. L. RUSSELL, *Mathematics of Finite Dimensional Control Systems, Theory and Design*, Marcel Dekker, New York, 1979.

[19] D. SALAMON, *Structure and stability of finite dimensional approximations for functional differential equations*, SIAM J. Control Optim., 23 (1985), pp. 928-951.

[20] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 97-112, January 1990

005

ON THE OPTIMAL REWARD FUNCTION OF THE CONTINUOUS TIME MULTIARMED BANDIT PROBLEM*

JOSÉ LUIS MENALDI† AND MAURICE ROBIN‡

Abstract. The optimal reward function associated with the so-called "multiarmed bandit problem" for general Markov–Feller processes is considered. It is shown that this optimal reward function has a simple expression (product form) in terms of individual stopping problems, without any smoothness properties of the optimal reward function neither for the global problem nor for the individual stopping problems. Some results relative to a related problem with switching cost are obtained.

Key words. variational inequality, switching problem, bandit problem, dynamic programming, index policy

AMS(MOS) subject classifications. 35B37, 49A60, 49B60, 60J25, 93E20

1. Introduction. This paper deals with the properties of the optimal reward function associated with the so-called "multiarmed bandit problem." Let us recall, formally, the statement of the problem: assume that there are N independent machines. $x_i(t)$, $t \in \mathbb{R}_+$ is the state (for instance the production) of machine i. At each time t, one operates only one machine, the others being frozen. When machine i is operating, $x_i(t)$ evolves as a continuous time Markov process with a given semigroup $\Phi^i(t)$. If $i(t)$ denotes the number of the machine in operation at time t, we want to maximize a global payoff

$$J = E \int_0^\infty e^{-\alpha t} f(i(t), x_{i(t)}(t))\, dt \tag{1.1}$$

where f is a given instantaneous reward.

The multiarmed bandit problem has been studied by Gittins [4] and Whittle [8] in the discrete time case, and more recently by Varaiya, Walrand, and Buyukkoc [7] in a more general setting. Karatzas [5] studied the continuous time case when $x_i(t)$ is a one-dimensional diffusion process. The most general study is done in Mandelbaum [13], [14] who formulated the problem as the control of a multiparameter process. This approach allows, in particular, a strong formulation of the optimal process when $x_i(t)$ is a diffusion process.

In Whittle [9] it is shown that the optimal reward function has a simple expression in terms of an individual stopping problem each involving only one machine. Such an expression is shown to hold true for the diffusion bandit problem in Karatzas [5] thanks to the smoothness of the reward function which allows explicit computations.

In this paper, the main objective is to obtain such an expression when the $x_i(t)$ are general Feller processes, without smoothness properties of the optimal reward function neither for the global problem nor for the individual stopping problem.

Let us describe briefly what expression we are looking for.

* Received by the editors October 19, 1987; accepted for publication (in revised form) March 9, 1989.

† Wayne State University, Department of Mathematics, Detroit, Michigan 48202. This research was partially supported by National Science Foundation grant DMS-8601998 and Air Force Office of Scientific Research contract F49620-86-C-0111.

‡ Institut National de Recherche en Informatique et en Automatique, Rocquencourt, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay, France.

Following Whittle [9], we will use the variant of the problem where one can decide, at any time, to stop the control problem, with a reward M if this "retirement option" is chosen.

Assume that $x_i(t)$ is for each i, a Markov process with values in some space E_i, with semigroup $\Phi^i(t)$.

If $\underline{x}$ denotes the initial state of the whole set of machines, and if $u(\underline{x}, M)$ is the corresponding optimal reward function, then by applying, formally, the dynamic programming arguments, $u(\underline{x}, M)$ is shown to be the minimum solution of the following inequalities:

$$u(\underline{x}, M) \geqq e^{-\alpha t}\Phi^i(t)u(\underline{x}, M) + \int_0^t e^{-\alpha s}\Phi^i(s)f_i(x_i)\,ds \tag{1.2}$$

$$u(\underline{x}, M) \geqq M.$$

The individual stopping problems have optimal cost functions $(\phi_i(x_i, M), i = 1, N)$, where ϕ_i is the minimum solution of

$$\phi_i(x_i, M) \geqq e^{-\alpha t}\Phi^i(t)\phi_i(x_i, M) + \int_0^t e^{-\alpha s}\Phi^i(s)f_i(x_i)\,ds \tag{1.3}$$

$$\phi_i(x_i, M) \geqq M$$

when $\alpha k \leqq f_i(x_i) \leqq \alpha K, \quad \forall i, \quad \forall x_i$.

The objective is to show that

$$u(\underline{x}, M) = K - \int_M^K \prod_{i=1}^N \frac{\partial \phi_i}{\partial m}\,dm. \tag{1.4}$$

It would be nice to obtain such a formula by analytic methods, as it can be shown that (1.2) and (1.3) have a minimal solution (cf. [1], [2], [3]). However, without smoothness on ϕ_i, we do not know how to show the result by analytic methods.

Here we will use an intermediary control problem (§ 2.1) which is suitable for our objective, although it does not contain a general statement of the multiarmed bandit itself when there is no switching cost.

Using this particular interpretation of the minimal solution of (1.2), we will show (1.4) using an extension to the continuous time case of the Tsitsiklis' lemma [6]. In § 3, we investigate the problem with switching cost, showing a similar lemma; it does not seem possible, however, to obtain an expression of the optimal reward in terms of some individual problems.

2. Problem without switching cost. We start with a control problem which will provide a stochastic interpretation of (1.2).

2.1. An intermediary control problem. Let E_i, $i = 1 \cdots N$ be a family of compact metric spaces endowed with their Borel σ-algebra.

Define $E = E_1 \times \cdots \times E_N$. Throughout the paper,

$\underline{x}$ will denote an element of E, i.e.,

$$\underline{x} = (x_1, \cdots, x_N), \qquad x_i \in E_i.$$

We are given a family of Markov semigroups $\Phi^i(t)$ $i = 1, \cdots, N$, $\Phi^i(t)$ being defined and *continuous* on $C(E_i)$, the Banach space of continuous functions on E_i.[1]

[1] So, Φ^i is a Feller semigroup on $C(E_i)$, cf. Dynkin [10].

If $\Omega_i = D(\mathbb{R}_+, E_i)$, the space of right continuous, left limited functions on $\mathbb{R}_+$ with values in E_i, we denote by $Q^i_{x_i}$ the probability measure on Ω_i corresponding to Φ^1, and we define

$$\Omega = \Omega_1 \times \cdots \times \Omega_N$$

and $\{F_t\}$ the associated canonical σ-algebra.

In order to define the controlled process, we first consider the probability measure corresponding to constant trajectories for the components $j \neq i$ (i being the number of the process which is active, the others being frozen), and which gives the markovian evolution corresponding to $\Phi^i(t)$ for the component i: in other words we define

$$P_{i,\underline{x}} = \delta_{x_1} \times \cdots \times \delta_{x_{i-1}} \times Q^i_{x_i} \times \delta_{x_{i+1}} \cdots \times \delta_{x_N}. \tag{2.1}$$

Notice that, if

$$\underline{x}_t(\omega) = \omega(t) \quad \text{for } \omega \in \Omega,$$

then

$$E_{i,\underline{x}} g(\underline{x}_t) = E^i_{x_i} g(x_1, \cdots, x_{i-1}, x_i(t), x_{i+1}, \cdots, x_N)$$

where $E_{i,\underline{x}}$ (respectively, $E^i_{x_i}$) denotes the expectation with respect to $P_{i,\underline{x}}$, (respectively, $Q^i_{x_i}$).

Assume now that

(2.2) $f_i(x_i)$ is a positive function $f_i \in C(E_i)$, $\forall i$ $\alpha > 0$ a discount factor

(2.3) V will be the set of admissible controls and $v \in V \Leftrightarrow v = (\theta_n, \xi_n)_{n \geqq 0}$, $\theta_0 = 0$, where (θ_n) is an increasing sequence of F_t stopping times, ξ_n a F_{θ_n}-measurable random variable with values in $\{1, \cdots, N\}$ and we assume

(2.4) $\theta_n(\omega) \uparrow +\infty \quad \forall \omega.$

For any $v \in V$, $\underline{x} \in E$, we define, as in [11], the following sequence of probability measures on (Ω, F_∞), if $\xi_0 = i$

$$P^0 = P_{i,\underline{x}}$$

P^1 is the (unique) probability measure on (Ω, F_∞) such that

$$P^1 = P^0 \text{ on } F_{\theta_1}$$

$$P^1(\eta_{\theta_1} B \,|\, F_{\theta_1}) = P_{\xi_1, \underline{x}_{\theta_1}}(B), \quad P^0 \text{ a.s.},$$

$\forall B$ Borel subset of Ω, η_t being the shift operator,

and so on $\cdots$

P^n is similarly defined from P^{n-1}

$$P^n = P^{n-1} \quad \text{on } F_{\theta_n}$$

$$P^n(\eta_{\theta_1} B \,|\, F_{\theta_n}) = P_{\xi_n, \underline{x}_{\theta_n}}(B), \quad P^{n-1} \text{ a.s.}$$

Defining

$$\xi(t) = \xi_n \quad \text{for } t \in [\theta_n, \theta_{n+1}[, \qquad n \geqq 0.$$

We consider the discounted reward

$$J_{\underline{x}}(v) = \lim_{n \uparrow \infty} E^n_{\underline{x}} \int_0^{\theta_{n+1}} e^{-\alpha t} f(\xi(t), \underline{x}_t)\, dt$$

where

$$f(\xi, \underline{x}) = f_i(x_i) \quad \text{iff } \xi = i.$$

Actually, with the assumptions $\theta_n \uparrow +\infty$, one can know that there exists a unique probability measure $P^v_{i,\underline{x}}$ on (Ω, F_∞) such that

$$P^v_{i,\underline{x}} = P^n \quad \text{on } F_{\theta_n} \tag{2.5}$$

and one can also define our total reward by

$$J_{\underline{x}}(v) = E^v_{\underline{x}} \int_0^\infty e^{-\alpha t} f(\xi_t, \underline{x}_t)\, dt. \tag{2.6}$$

We now add another control possibility, namely the "retirement option."

Let T be the set of F_t stopping times, for $v \in V$, $\tau \in T$, and $(i, \underline{x}) \in U \times E$, we define the total reward as

$$J^M_{\underline{x}}(v, \tau) = E^v_{\underline{x}} \left\{ \int_0^\tau e^{-\alpha t} f(\xi_t, \underline{x}_t)\, dt + e^{-\alpha \tau} M \right\} \tag{2.7}$$

where M is a given constant.

We will use, as in Whittle [9], the additional assumption

$$\alpha k \leqq f_j \leqq \alpha K, \quad \forall j \in U, \tag{2.8}$$

where $k < K$ are given nonnegative constants.

The optimal reward function is

$$u(\underline{x}, M) = \operatorname{Sup} (J^M_{\underline{x}}(v, \tau), (v, \tau) \in V \times T). \tag{2.9}$$

Using a *formal* dynamic programming argument, it is easy to check that $u(\underline{x}, M)$ should solve the following inequalities

$$w(\underline{x}, M) \geqq e^{-\alpha t} \Phi^i(t) w + \int_0^t e^{-\alpha s} \Phi^i(s) f_i(x_i)\, ds, \quad \forall t > 0 \quad \forall i \in U,$$

$$w(\underline{x}, M) \geqq M, \tag{2.10}$$

$$w(\cdot, M) \text{ is a bounded measurable function.}$$

In the following section, we will show that u is actually the *minimum* solution of these inequalities (for fixed M).

Let us recall the following result (cf. Bensoussan and Robin [3], Bensoussan [1]):

THEOREM 2.1. *Under the assumption* (2.2) *there exists a minimum solution* $\bar{u}$ *of* (2.10) *in the space of bounded measurable functions. Moreover* $\bar{u}$ *is upper semicontinuous.*

Remark 2.1. In Bensoussan and Robin [3], another kind of interpretation was given for $\bar{u}(\underline{x}, M)$. The present one will be more suitable for the problem we consider.

2.2. Characterization of the optimal reward (2.9). In order to characterize $u(\underline{x}, M)$ as defined in (2.9), we introduce another switching problem, with a switching cost ε. Namely, we consider the same problem as in § 2.1, but now, at each switching time a cost ε (i.e., a reward $-\varepsilon$) is involved. This is in fact a classical switching problem (which can be considered as an impulse control problem where the state is $(\xi_t, \underline{x}_t)$, cf. Bensoussan [1], Bensoussan and Lions [2] for the general theory).

In this context, let

$$V_0 = \{v = (\theta_n, \xi_n)_{n \geqq 1}\}$$

be the set of admissible controls, θ_n, ξ_n being defined as previously.

For $(i, \underline{x}) \in U \times E$, define the reward

$$J_{i,\underline{x}}^{M,\varepsilon}(v, \tau) = E_{i,\underline{x}}^{v}\left\{\int_0^{\tau} e^{-\alpha s} f(\xi(t), \underline{x}_t)\, dt - \varepsilon \sum_{j \geqq 1} e^{-\alpha \theta_j} \chi_{\theta_j < \tau} + e^{-\alpha \tau} M\right\} \tag{2.11}$$

where $E_{i,\underline{x}}^{v}$ is defined as in § 2.1, $(v, \tau) \in V_0 \times T$, and $\chi_B(\omega)$ is the characteristic function of the set B and $\xi_0 = i$ for the construction of $P_{i,\underline{x}}^{n}$.

We also define

$$u_i^{\varepsilon}(\underline{x}, M) = \sup\,(J_{i,\underline{x}}^{M,\varepsilon}(v, \tau), (v, \tau) \in V_0 \times T). \tag{2.12}$$

Let $u^{\varepsilon} = (u_1^{\varepsilon}, \cdots, u_N^{\varepsilon})$.

From impulse control theory (cf. Bensoussan and Lions [2], [11]) we know that, for fixed M, u^{ε} is the minimum element of the set of bounded measurable functions w satisfying

$$\begin{aligned} &w_i(\underline{x}) \geqq e^{-\alpha t}\Phi^i(t) w_i + \int_0^t e^{-\alpha s}\Phi^i(s) f_i(x_i)\, ds, \quad \forall t > 0, \\ &w_i(\underline{x}) \geqq -\varepsilon + \max_j w_j(\underline{x}), \\ &w_i(\underline{x}) \geqq M. \end{aligned} \tag{2.13}$$

Moreover, $u_i^{\varepsilon}(\underline{x}) \in C(E)$, $\forall i = 1, \cdots, N$.

We first establish the following result.

THEOREM 2.2. *Let $\underline{u}(\underline{x}, M)$ be the minimum solution of the inequalities* (2.10), *then*

$$\lim_{\varepsilon \downarrow 0} u_i^{\varepsilon}(\underline{x}, M) = \underline{u}(\underline{x}, M) \tag{2.14}$$

pointwise in $\underline{x}$.

Proof. It is clear that $u_i^{\varepsilon}(\underline{x}, M)$ increases when ε decreases, and that $u_i^{\varepsilon}(\underline{x}, M)$ is bounded (say by $(1/\alpha)\|f\| + M$). Let us define

$$\underline{w}_i = \lim_{\varepsilon \downarrow 0} u_i^{\varepsilon}.$$

From (2.13), we have

$$\underline{w}_i \geqq \max_j w_j(\underline{x}, M), \quad \forall i. \tag{2.15}$$

Hence

$$\underline{w}_i(\underline{x}, M) = \underline{w}(\underline{x}, M) \quad \forall i.$$

But $\underline{u}(\underline{x}, M)$, the minimum solution of (2.10), satisfies obviously (2.13) and therefore

$$u_i^{\varepsilon}(\underline{x}, M) \leqq \underline{u}(\underline{x}, M) \quad \forall \varepsilon, i.$$

So we deduce, when $\varepsilon \to 0$,

$$\underline{w}(\underline{x}, M) \leqq \underline{u}(\underline{x}, M). \tag{2.16}$$

But we see that $\underline{w}(\underline{x}, M)$ will also satisfy (2.10), since this is identical to (2.13) when $\varepsilon = 0$ for a function which does not depend explicitly on i.

Therefore

$$\underline{w}(\underline{x}, M) \geqq \underline{u}(\underline{x}, M).$$

Hence

$$\underline{w}(\underline{x}, M) = \underline{u}(\underline{x}, M)$$

and the theorem is proved. □

Let us define

$$u(\underline{x}, M) = \sup (J_{\underline{x}}^{M}(v, \tau), (v, \tau) \in V \times T). \tag{2.17}$$

Then we have the following Theorem.

THEOREM 2.3.

$$\underline{u}(\underline{x}, M) = u(\underline{x}, M). \tag{2.18}$$

Proof. Since $\varepsilon > 0$, we have

$$u_i^{\varepsilon}(\underline{x}, M) = \sup (J_{i,\underline{x}}^{M,\varepsilon}(v, \tau), (v, \tau) \in V_0 \times T)$$

and

$$J_{i,\underline{x}}^{M,\varepsilon}(v, \tau) \leqq J_{\underline{x}}^{M}(\tilde{v}, \tau)$$

where $\tilde{v} = ((i, 0), v)$, and $(\tilde{v}, \tau) \in V \times T$.

Therefore

$$u_i^{\varepsilon}(\underline{x}, M) \leqq u(\underline{x}, M),$$

hence

$$\underline{u}(\underline{x}, M) \leqq u(\underline{x}, M). \tag{2.19}$$

Now, for any solution w of the inequalities (2.10), one can show as in [11, Thm. VII, § 3.1] or [2b, § 6.4], that

$$w(\underline{x}) \geqq E_x^m \left\{ e^{-\alpha\theta_{m+1}\wedge\tau} w(\underline{x}_{\theta_{m+1}\wedge\tau}) + \int_0^{\theta_{m+1}\wedge\tau} e^{-\alpha t} f(\underline{x}_t, v_t)\, dt \right\}$$

for any admissible control (v, τ), $v = (\theta_i, \xi_i)_{i \geqq 0}$ where E_x^m is the expectation corresponding to the measure P_x^m associate to (v, τ) as in § 1, with $\theta_m \wedge \tau$ instead of θ_m.

From this inequality, we deduce, when $m \to +\infty$, since $w(\underline{x}) \geqq M$ and $\theta_m \wedge \tau \uparrow \tau$, that

$$w(\underline{x}) \geqq J_{\underline{x}}^{M}(v, \tau)$$

and therefore

$$w(\underline{x}) \geqq u(\underline{x}, M).$$

Finally, this gives

$$\underline{u}(\underline{x}, M) \geqq u(\underline{x}, M),$$

which, with (2.19), proves the result. □

2.3. Reduction to write off policies. Following Whittle, a write off policy is defined as a policy such that there exists a family of "write-off" sets $S_i \subset E_i$ with the following properties.

- as soon as x_i (the state of the process i) belongs to S_i, the process i is abandoned;
- one retires as soon as all the processes have been abandoned, and only then;
- before retiring, one works only with those processes which have not been abandoned.

In this section we are going to show a lemma similar to the one obtained by Tsitsiklis [6] for discrete time, showing that we can restrict ourselves to write off policies with write off sets defined by optimal stopping problems for the individual processes.

The individual stopping problems. Let us consider the optimal stopping reward

$$\phi_i(x_i, M) = \sup_{\tau} I^M_{x_i}(\tau) \tag{2.20}$$

$$I^M_{x_i}(\tau) = E^i_{x_i}\left[\int_0^\tau e^{-\alpha t} f_i(x_{it})\, dt + e^{-\alpha\tau} M\right]. \tag{2.21}$$

It is known from standard theory (see Bensoussan [1]) that $\phi_i(x_i, M)$ is the minimum element of the set of functions $w(x_i)$ satisfying

$$\begin{aligned} &w(x_i) \geqq e^{-\alpha t}\Phi^i(t) w + \int_0^t e^{-\alpha s}\Phi^i(s) f_i(x_i)\, ds, \quad \forall t > 0 \\ &w(x_i) \geqq M, \qquad w \in C(E_i). \end{aligned} \tag{2.22}$$

Let us show the following results which extend the discrete time case (cf. Whittle [9]) and the diffusion case (Karatzas [5]).

LEMMA 2.1. *Under the assumptions* (2.2)-(2.8), $\phi(x, M) = \phi_i(x_i, M)$ *has the following properties*:

(i) $\phi(x, M) = M \ \forall M \geqq K$;

(ii) $\phi(x, M) = \int_0^\infty e^{-\alpha t}\Phi^i(t) f_i(x)\, dt, \ \forall M \leqq k$;

(iii) $\forall x \in E_i$, $\phi(x, \cdot)$ *is an increasing convex function*;

(iv) $\phi(x, \cdot)$ *is Lipschitz continuous and in every* M *where the derivative exists*

$$0 \leqq \frac{\partial \phi}{\partial M} \leqq 1;$$

(v) *in every point where the derivative exists*

$$\frac{\partial \phi}{\partial M}(x, M) = E_x\, e^{-\alpha\hat{\tau}}$$

where $\hat{\tau}$ *is optimal for* (2.20), *namely*

$$\hat{\tau} = \inf(t \geqq 0, \phi(x, M) = M).$$

Proof. (i) This shows that $\tau = 0$ is optimal in (2.20) whenever $M \geqq K$. Since $f_i \leqq \alpha K$

$$\begin{aligned} J^M_x(\tau) &\leqq E_x\{(1 - e^{-\alpha\tau})K + e^{-\alpha\tau} M\} \\ &= K + E_x\, e^{-\alpha\tau}(M - K), \end{aligned}$$

clearly if $M \geqq K$, $\tau = 0$ gives the maximum value.

(ii) $f_i \geqq \alpha k$ implies

$$w_o(x) = \int_0^\infty e^{-\alpha t}\Phi^i(t) f_i\, dt \geqq k$$

and since $w_0(x) = e^{-\alpha t}\Phi^i(t) w_0 + \int_0^t e^{-\alpha s}\Phi^i(s) f_i\, ds$, we see that w_0 satisfies (2.22) for $M = k$.

Moreover $t = +\infty$ in (2.22) gives

$$w(x) \geqq w_0(x) \quad \forall w \text{ solution of (2.22)}.$$

(iii) If $0 \leqq \lambda \leqq 1$, then we check that

$$w_\lambda = \lambda\phi(x, m_1) + (1 - \lambda)\phi(x, m_2)$$

satisfies (2.22) for $m = \lambda m_1 + (1-\lambda)m_2$ and therefore

$$w_\lambda \geqq \phi(x, \lambda m_1 + (1-\lambda)m_2).$$

The increasing property is obvious from (2.20).

(iv) From (2.21) one has, for an arbitrary τ

$$I_x^{M+\delta}(\tau) - I_x^M(\tau) = E_x e^{-\alpha\tau}\delta$$

therefore, for $\delta > 0$

$$I_x^{M+\delta}(\tau) \leqq I_x^M(\tau) + \delta \leqq \phi(x, M) + \delta$$

implying

$$\phi(x, M+\delta) \leqq \phi(x, M) + \delta$$

and since $\phi(x, M+\delta) \geqq \phi(x, M)$, we see that

$$0 \leqq \frac{\partial^+\phi}{\partial M}(x, M) \leqq 1.$$

(v) Let $\hat{\tau} = \inf(t \geqq 0, \phi(x, M) = M)$, we know (cf. [1]) that

$$\phi(x, M) = I_x^M(\hat{\tau}).$$

Therefore, if $\delta > 0$,

$$\begin{aligned} I_x^{M+\delta}(\hat{\tau}) &= I_x^M(\hat{\tau}) + \delta E_x e^{-\alpha\hat{\tau}} \\ &= \phi(x, M) + \delta E_x e^{-\alpha\hat{\tau}} \end{aligned}$$

hence

$$\phi(x, M+\delta) - \phi(x, M) \geqq \delta E_x e^{-\alpha\hat{\tau}}$$

$$\frac{\partial^+\phi}{\partial M}(x, M) \geqq E_x e^{-\alpha\hat{\tau}}.$$

Taking $\delta < 0$, we get

$$\frac{\partial^-\phi}{\partial M}(x, M) \leqq E_x e^{-\alpha\hat{\tau}}.$$

Therefore, in M such that the derivative exists, we get the result. □

COROLLARY. *$\partial^+\phi/\partial M$ is a right continuous increasing function such that*

$$0 \leqq \frac{\partial^+\phi}{\partial M} \leqq 1,$$

$$\frac{\partial^+\phi}{\partial M}(x, M) = 1 \quad \forall M \geqq K$$

$$\frac{\partial^+\phi}{\partial M}(x, M) = 0 \quad \forall M < k.$$

Let us now define, for fixed i

$$y_i = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_N)$$

$U_i(y_i, M)$ the optimal reward function when only the processes different from i are available.

From the previous section, $U_i(y_i, M)$ is the minimum solution of

$$W_i(y_i, M) \geqq e^{-\alpha t}\Phi^j(t)\, W_i + \int_0^t e^{-\alpha s}\Phi^j(s) f_j(x_j)\, ds, \quad \forall t > 0, \quad \forall j \neq i \tag{2.23}$$

$$W_i(y_i, M) \geqq M, \quad W_i(\cdot, M) \text{ bounded and measurable.}$$

We can now state the Tsitsiklis' lemma in continuous time.

LEMMA 2.2. (*Tsitsiklis' lemma in continuous time*). *One has*

$$u(\underline{x}, M) \leqq \phi_i(x_i, M) - M + U_i(y_i, M), \quad \forall i. \tag{2.24}$$

Proof. Let $W_i(\underline{x}, M)$ be the right-hand side of (2.24). We are going to show that W_i satisfies (2.10) and since u is the minimum solution, this will show the lemma. Notice that since $U_i \geqq M$ and $\phi_i \geqq M$, we have

$$W_i(\underline{x}, M) \geqq M.$$

Moreover, since U_i does not depend on x_i,

$$e^{-\alpha t}\Phi^i(t)\, W_i + \int_0^t e^{-\alpha s}\Phi^i(s) f_i\, ds = \left\{ e^{-\alpha t}\Phi^i(t)\phi_i + \int_0^t e^{-\alpha s}\Phi^i(s) f_i\, ds \right\} + e^{-\alpha t}[U_i - M]$$

$$= I + II.$$

We have,

$$I \leqq \phi_i \quad \text{by (2.22)}$$

and, since $U_i - M \geqq 0$,

$$II \leqq U_i - M.$$

Then, for $j \neq i$, since ϕ_i does not depend on x_j, $j \neq i$,

$$e^{-\alpha}\Phi^j(t)\, W_i + \int_0^t e^{-\alpha s}\Phi^j(s) f_j\, ds$$

$$= \left\{ e^{-\alpha t}\Phi^j(t)\, U_i + \int_0^t e^{-\alpha s}\phi^j(s) f_j\, ds \right\} + e^{-\alpha t}[\phi_i - M] = III + IV.$$

Hence, using (2.23)

$$III \leqq U_i,$$

and

$$IV \leqq \phi_i - M \text{ since } \phi_i - M \geqq 0.$$

Therefore the lemma is proved. □

COROLLARY. *Define* $S_i^M = \{x_i \in E_i, \phi_i(x_i, M) = M\}$, *then one can restrict the policies to be write off with respect to* $(S_i^M, i = 1, \cdots, N)$.

Proof. Notice that $U_i(y_i, M) \leqq u(\underline{x}, M)$. If $x_i \in S_i^M$, then (2.24) gives, with the above inequality,

$$u(\underline{x}, M) = U_i(y_i, M)$$

which means that the optimal reward is the same as the one with $N-1$ processes where the i process has been dropped. If there exists i such that, for $\underline{x} = (x_1, \cdots, x_N)$, $x_i \notin S_i^M$, then $u(\underline{x}, M) \geqq \phi_i(x_i, M) > M$ implies that it is not optimal to retire. Finally

if $\underline{x}$ is such that $x_i \in S_i^M$, $\forall i$, then $u(\underline{x}, M) = U_i$ $\forall i$ and we can use the same argument for N-1 processes to show that

$$u(\underline{x}, M) = \phi_j(x_j, M) = M \quad \forall j. \qquad \square$$

Let us denote by V_{off}^M the set of admissible write off policies corresponding to $(S_i^M,\ i = 1, \cdots, N)$. We will use the following lemma due to Whittle (cf. [9]).

LEMMA 2.3. *If (v, τ) is a write off policy, then*

$$E_{\underline{x}}^v e^{-\alpha\tau} = \prod_{i=1}^{N} E_{x_i} e^{-\alpha\tau_i}$$

where τ_i is the retirement times when only the process i is available.

Proof. For the proof see Whittle [9]. $\square$

In our context, this means that, if

$$\tau_i^M = \inf(t \geqq 0,\ \phi_i(x_{it}, M) = M), \tag{2.25}$$

then, for all write off policies,

$$E_x^v e^{-\alpha\tau} = \prod_i E_{x_i} e^{-\alpha\tau_i^M}. \tag{2.26}$$

We can then deduce the product formula for u.

THEOREM 2.4.

$$U(\underline{x}, M) = K - \int_M^K \sum_{i=1}^{N} \frac{\partial \phi_i}{\partial m}(x_i, m)\, dm. \tag{2.27}$$

Proof. Let (v, τ) be an admissible write off policy with respect to $(S_i^M,\ i = 1, \cdots, N)$. We have,

$$J_{\underline{x}}^{M+\delta}(v, \tau) - J_{\underline{x}}^M(v, \tau) = \delta E_{\underline{x}}^v e^{-\alpha\tau}.$$

From the previous lemma

$$J_{\underline{x}}^{M+\delta}(v, \tau) - J_{\underline{x}}^M(v, \tau) = \delta \cdot \prod_{i=1}^{N} E_{x_i} e^{-\alpha\tau_i^M},$$

therefore

$$u(\underline{x}, M+\delta) \geqq J_{\underline{x}}^M(v, \tau) + \delta \prod_{i=1}^{N} E_{x_i} e^{-\alpha\tau_i^M}.$$

Note that the last term is independent from (v, τ) as far as (v, τ) is a write off policy with respect to (S_i^M). Therefore, maximizing with respect to (v, τ)

$$u(\underline{x}, M+\delta) \geqq u(\underline{x}, M) + \delta \prod_i E_{x_i} e^{-\alpha\tau_i^M}$$

which implies, for $\delta > 0$,

$$\frac{\partial^+ u}{\partial M}(\underline{x}, M) \geqq \prod_i E_{x_i} e^{-\alpha\tau_i^M}$$

and for $\delta < 0$

$$\frac{\partial^- u}{\partial M}(\underline{x}, M) \leqq \prod_i E_{x_i} e^{-\alpha\tau_i^M}.$$

Therefore, at every point where the derivatives exist, we have, thanks to the Lemma 2.1,

$$\frac{\partial u}{\partial M}(\underline{x}, M) = \prod_{i=1}^{N} \frac{\partial \phi_i}{\partial M}(x_i, M).$$

Integrating from M to K, using the fact that $u(\underline{x}, K) = K$, we get (2.27). □

Remark. From Bensoussan and Robin [3], we can show that the optimal reward of the discrete time problem converges to the $u(\underline{x}, M)$ when the time step h goes to zero. However, we have not been able to show the product formula in continuous time by letting h go to zero on the product formula of the discrete time case.

2.4. The forward induction lemma. Let us consider the discrete time version of the stopping problems (2.22). Namely, for $h > 0$, we define (dropping the index i)

$$r_h(x) = E_x \int_0^h e^{-\alpha s} f(x_s)\, ds$$

$$Q_h z = \phi(h) z$$

$$\beta = e^{-\alpha h}.$$

Then the optimal reward for the discrete stopping problem $\phi_h(x, m)$ is the unique solution of

$$\phi_h(x, m) = \max\,(r_h + \beta Q_h \phi_h, m).$$

Defining $V_h = \{\tau$, stopping times with values in $N_h = \{nh, n \geqq 0\}\}$, we can write

$$\phi_h(x, m) = \sup_{\tau \in V_h} E_x \left(\int_0^\tau e^{-\alpha s} f(x_s)\, ds + e^{-\alpha \tau} m \right).$$

The *index* is defined, as previously, as

$$M_h(x) = \inf\,(m > k,\ \phi_h(x, m) = m).$$

On the other hand, Whittle [9] shows that $M_h(x)$ has the following representation:

$$M_h(x) = \sup_{\tau \in V_h^*} \frac{E_x \int_0^\tau e^{-\alpha s} f(x_s)\, ds}{1 - E_x e^{-\alpha \tau}} \tag{2.28}$$

with $V_h^* = \{\tau$ stopping times with values in $N_h^* = N_h - \{0\}\}$.

The extension of the formula (2.28) to diffusion processes was done by Karatzas [5] using explicit calculation for one-dimensional processes. We are going to show the same formula in our context; the idea being to approximate the stopping problem (2.22) by a discrete time problem (like in Bensoussan–Robin [3]).

LEMMA 2.4. *Let $\phi(x, m)$ be defined as in* (2.20) (*where we drop the index i*), *and define*

$$M(x) = \inf\,(m > k,\ \phi(x, m) = m), \text{ and } V^* = \bigcup_h V_h^*,$$

then

$$M(x) = \sup_{\tau \in V^*} \frac{E_x \int_0^\tau e^{-\alpha s} f(x_s)\, ds}{1 - E_x e^{-\alpha \tau}}$$

where $V^* = \{\bigcup_h V_h^*\}$.

Proof. Starting with $M_h(x)$ we have

$$k \leqq M_h(x) \leqq K \quad \forall x, \quad \forall h.$$

Clearly, V_h^* is increasing as h decreases to zero and therefore $M_h(x)$ is increasing when $h \downarrow 0$. For fixed x, let

$$v(x) = \lim_{h \downarrow 0} M_h(x)$$

then

$$v(x) = \sup_{\tau \in V^*} Z(\tau), \quad \text{where } Z(\tau) = \frac{E_x \int_0^\tau e^{-\alpha s} f(x_s)\, ds}{1 - E_x e^{-\alpha \tau}}.$$

Indeed, for all ε, there exists h, such that

$$v \geqq M_h > v - \varepsilon \Rightarrow \exists \delta(\varepsilon) \text{ s.t.} \quad M_h - \delta(\varepsilon) > v - \varepsilon$$

and from the definition of M_h, we can find $\tau_h(\delta(\varepsilon))$ such that, $\tau_h \in V_h^*$,

$$M_h \geqq Z(\tau_h) > M_h - \delta(\varepsilon).$$

Therefore for all ε, there exists $\tau \in V^*$ such that $v \geqq Z(\tau) > v - \varepsilon$ proving that $v = \sup (Z(\tau), \tau \in V^*)$.

Let us prove that $v = M(x)$. Assume that $m \geqq v$, then

$$m \geqq v \geqq \frac{E_x \int_0^\tau e^{-\alpha s} f(x_s)\, ds}{1 - E_x e^{-\alpha \tau}} \quad \forall \tau \in V^*$$

$\Rightarrow$

$$m \geqq E_x \left(\int_0^\tau e^{-\alpha s} f(x_s) + e^{-\alpha \tau} m \right) \quad \forall \tau \in V^*$$

(and for $\tau = 0$, we have the equality). Therefore $m \geqq \phi(x, m)$.

But $\phi(x, m) \geqq m$, for all m implies $\phi(x, m) = m$ for all $m \geqq v$. Now assume that $m < v$.

Let us assume that for such m

$$\phi(x, m) = \sup_\tau E_x \int_0^\tau e^{-\alpha s} f(x_s) + e^{-\alpha \tau} m = m.$$

This would imply

$$m \geqq \frac{E_x \int_0^\tau e^{-\alpha s} f(x_s)\, ds}{1 - E_x e^{-\alpha \tau}} \quad \forall \tau \in V^*$$

which contradicts the assumption $m < v$.

Therefore

$$m < v \Rightarrow \phi(x, m) > m, \quad \text{hence } v = M(x). \qquad \square$$

Remark. As it was stressed in Katehakis-Veinott [12] we can also characterize $M_h(x)$ using the "restart in x-problem" for which the optimal reward function $v_h^x(\cdot)$ is given by

$$v_h^x(y) = \max (r_h(y) + \beta Q_h v_h^x, r_h(x) + \beta Q_h v_h)$$

and then (see [12]) we have

$$M_h(x) = v_h^x(x).$$

In continuous time, in order to define a similar problem, we can use the discrete time solution; namely, as in Bensoussan-Robin [3], we could show that for $h = 2^{-N}$

$$v_N^x(y) = v_h^x(y)$$

is increasing when $N \to +\infty$, (and bounded).

Then

$$v^x(y) = \lim_{N \uparrow \infty} v_N^x(y)$$

is the minimum solution of the inequalities

$$v^x(y) \geqq e^{-\alpha t}\phi(t)v^x + \int_0^t e^{-\alpha s}\Phi(s)f\,ds$$

$$v^x(y) \geqq v^x(x), \quad \forall y$$

$v^x(x)$ bounded measurable functions.

This is the continuous time version of the restart in x-problem.

3. The problem with switching cost. We now turn back to the case where there is a switching cost incurred at each time we change the active process. This was already considered in § 2.2 when we constructed the functions u_i^ε. Recall that this is a more or less standard impulse control problem where the underlying state is in fact $(z, \underline{x})$ where $z \in \{1, \cdots, N\}$ is the number of the active process. It would be interesting to know if a product formula like (1.4) holds. We do not know the answer, neither for the question of the optimality of some index rule. However, we can show that the concept of write off policy is still valid in this case and this gives some more information on the optimal policies than the mere interpretation of the dynamic programming condition. The reduction to write off policies will be a consequence of the following simple result, similar to the Tsitsiklis' lemma. Let us make precise some notations: we drop the ε in the optimal reward which is now

$$u(z, \underline{x}, M) = \sup\,(J_{z,\underline{x}}^M(v, \tau), (v, \tau) \in V_0 \times T) \tag{3.1}$$

$J_{z,\underline{x}}^M(v, \tau)$, V_0, T being defined as in (2.11), with $z \in \{1, \cdots, N\}$. We know that u is the minimum element of the set of bounded and measurable functions $w(z, \underline{x})$ satisfying

$$w(z, \underline{x}) \geqq e^{-\alpha t}\Phi^z(t)w + \int_0^t e^{-\alpha s}\Phi^z(s)f_z(x_z)\,ds$$

$$w(z, \underline{x}) \geqq -\varepsilon + \max_j w(j, \underline{x}), \tag{3.2}$$

$$w(z, \underline{x}) \geqq M, \quad \forall z \in \{1, \cdots, N\}.$$

We denote by

$$y_i = (x_j, j \neq i),$$

$U(z, y_i, M)$ the optimal reward when only the processes different from i are available, and when the initial active process is the process number z.

LEMMA 3.1. *We have for arbitrary* $i \in \{1, \cdots, N\}$,

$$u(j, \underline{x}, M) \leqq [\phi_i(x_i, M) - (M + \varepsilon)]^+ + U(j, y_i, M) \quad \forall j \neq i \tag{3.3}$$

$$u(i, \underline{x}, M) \leqq \phi_i(x_i, M) - M + \max\left[M, -\varepsilon + \max_{j \neq i} U(j, y_i, M)\right]. \tag{3.4}$$

Proof. Let us define, for fixed i:

$$w(z,\underline{x})=\begin{cases}[\phi_i(x_i,M)-(M+\varepsilon)]^+ + U(z,y_i,M) & \text{for } z\neq i\\ \phi_i(x_i,M)-M+\max\left[M,-\varepsilon+\max\limits_{j\neq i}U(j,y_i,M)\right] & \text{for } z=i.\end{cases}$$

We are going to show that $w(z,\underline{x})$ satisfies (3.2) and since $u(z,\underline{x},M)$ is the minimum solution, this will prove the lemma. We have,

$$\begin{aligned}&e^{-\alpha t}\Phi^z(t)w(z,\underline{x})+\int_0^t e^{-\alpha s}\Phi^z(s)f_z(x_z)\,ds\\ &=\left\{e^{-\alpha t}\Phi^z(t)U(z,y_i,M)+\int_0^t e^{-\alpha s}\Phi^z(s)f_z(x_z)\,ds\right\}\\ &\quad+e^{-\alpha t}[\phi_i(x_i,M)-(M+\varepsilon)]^+ \quad \text{if } z\neq i\\ &=\left\{e^{-\alpha t}\Phi^i(t)\phi_i+\int_0^t e^{-\alpha s}\Phi^i(s)f_i(x_i)\,ds\right\}\\ &\quad+e^{-\alpha t}\left[\max\left[M,-\varepsilon+\max_{j\neq i}U(j,y_i,M)\right]-M\right] \quad \text{if } z=i.\end{aligned}$$

In the first case, thanks to (3.2), the right-hand side is less than

$$U(z,y_i,M)+e^{-\alpha t}[\phi_i(x_i,M)-(M+\varepsilon)]^+$$

i.e., less than

$$U(z,y_i,M)+[\phi_i(x_i,M)-(M+\varepsilon)]^+=w(z,\underline{x}).$$

In the second one, thanks to (2.22) for ϕ_i, the right-hand side is less than

$$\phi_i(x_i,M)-M+\max\left[M,-\varepsilon+\max_{j\neq i}U(j,y_i,M)\right]=w(z,\underline{x}) \quad \text{if } z=i.$$

Therefore the first inequality of (3.2) is satisfied. It is obvious that $w(z,\underline{x})\geqq M$. Now, for the second inequality of (3.2), we must check that

$$w(i,\underline{x})\geqq -\varepsilon+\max_{j\neq i}([\phi_i(x_i,M)-(M+\varepsilon)]^+ + U(j,y_i,M)) \tag{3.5}$$

and, for $z\neq i$

$$w(z,\underline{x})\geqq -\varepsilon+\max\begin{cases}[\phi_i(x_i,M)-(M+\varepsilon)]^+ + U(j,y_i,M) & \forall j\neq i\\ \phi_i(x_i,M)-M+\max\left[M,-\varepsilon+\max\limits_{j\neq i}U(j,y_i,M)\right].\end{cases} \tag{3.6}$$

But, since $\phi_i(x_i,M)-M\geqq[\phi_i(x_i,M)-(M+\varepsilon)]^+$, (3.5) is obvious from the definition of $w(i,\underline{x})$. For (3.6), since

$$U(z,y_i,M)\geqq -\varepsilon+\max_{j\neq i}U(j,y_i,M)$$

we have

$$w(z,\underline{x})\geqq -\varepsilon+[\phi_i(x_i,M)-(M+\varepsilon)]^+ +\max_{j\neq i}U(j,y_i,M)$$

and since $[\phi_i(x_i, M)-(M+\varepsilon)]^+ \geqq \phi_i(x_i, M)-M-\varepsilon$, we also have

$$w(z, \underline{x}) \geqq -\varepsilon + \phi_i(x_i, M) - M + \max\left[M, -\varepsilon + \max_{j \neq i} U(j, y_i, M)\right].$$

Therefore, the lemma is obtained. □

Let us define the following write off sets:

$$S_i^M = \{(z, \underline{x}) \in \{1, \cdots, N\} \times E_i \text{ such that either } z = i \text{ and } \phi_i(x_i, M) = M, \text{ or } z \neq i \text{ and } \phi_i(x_i, M) \leqq M + \varepsilon\}. \tag{3.7}$$

THEOREM 3.1. *We can restrict the admissible policies to be write off with respect to* $(S_i^M, i = 1, \cdots, N)$, *in other words*

(i) *if* $\exists i$ *s.t.* $(z, \underline{x}) \notin S_i^M$, *we continue* (*i.e., we do not use the retirement option*)

(ii) *if* $\forall i$, $(z, \underline{x}) \in S_i^M$, *we retire*

(iii) *if* $(z, \underline{x}) \in S_i^M$, *the process i is abandoned.*

Proof. (i) Assume that $\exists i$ s.t. $(z, \underline{x}) \notin S_i^M$

then -either $z = i$ and $\phi_i(x_i, M) > M$ hence $u(i, \underline{x}, M) \geqq \phi_i(x_i, M) > M$,

therefore we do not retire;

-or $z \neq i$ and $\phi_i(x_i, M) > M + \varepsilon$

hence $u(z, \underline{x}, M) \geqq -\varepsilon + \max_j u(j, \underline{x}, M) \geqq -\varepsilon + \phi_i > M$.

Therefore we do not retire.

(ii) Assume that

$$\forall i, \quad (z, \underline{x}) \in S_i^M \tag{3.8}$$

and to fix the idea, take $z = N$, then (3.3) implies, since $(z, \underline{x}) \in S_1^M$, and $U(z, y_1, M) \leqq u(z, \underline{x}, M)$,

$$u(z, \underline{x}, M) = U(z, y_1, M).$$

Denote U by $U^{N-1}(z, y_1^{N-1}, M)$ to make explicit that U is the optimal reward of a problem where only the $N-1$ first components are available, i.e., $y_1^{N-1} = (x_2, \cdots, x_N)$.

Then applying again (3.3) to the $N-1$ dimensional bandit problem we get, with $i = 2$

$$u(z, \underline{x}, M) = U^{N-1}(z, y_1^{N-1}, M) = U^{N-2}(z, y_2^{N-2}, M)$$

with $y_2^{N-1} = (x_3, \cdots, x_N)$.

This process goes on until

$$u(z, \underline{x}, M) = U^1(z, x_z, M) = \phi_z(x_z, M)$$

which by the assumption and (3.4) is equal to M. Therefore we must retire if (3.8) holds.

(iii) Assume $(z, \underline{x}) \in S_i^M$

-either $z \neq i$ then (3.3) and $u(z, \underline{x}, M) \geqq U(z, y_i, M)$ implies $u(z, \underline{x}, M) = U(z, y_i, M)$ meaning that we never use again the process i

-or $z = i$ and $\phi_i(x_i, M) = M$, then (3.4) implies that either we retire, or we have

$$u(z, \underline{x}, M) = -\varepsilon + \max_{j \neq z} U(j, y_z, M)$$

meaning that we switch to another process and never use the process $z = i$. This completes the proof of Theorem 3.1. □

Acknowledgment. The authors would like to thank Professors A. Bensoussan and P. L. Chow for the useful discussions on this work.

REFERENCES

[1] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, 1982.
[2a] A. BENSOUSSAN AND J. L. LIONS, *Inéquations variationnelles et problèmes d'arrêt optimal*, Dunod, Paris, 1978.
[2b] ———, *Applications des inéquations quasi variationnelles au contrôle stochastique*, Dunod, Paris, 1982.
[3] A. BENSOUSSAN AND M. ROBIN, *On the convergence of the discrete time dynamic programming equation for general semi-group*, SIAM J. Control Optim., 20 (1982), pp. 722-746.
[4] J. C. GITTINS, *Bandit processes and dynamic allocation indices*, J. Roy. Stat. Soc., 41 (1979), pp. 148-177.
[5] I. KARATZAS, *Gittins indices in the dynamic allocation problem for diffusion processes*, Ann. Probab., 12 (1984), pp. 173-192.
[6] J. N. TSITSIKLIS, *A lemma on the multiarmed bandit problem*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 576-577.
[7] P. VARAIYA, J. WALRAND, AND C. BUYUKKOC, *Extensions of the multiarmed bandit problem: the discounted case*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 426-439.
[8] P. WHITTLE, *Multiarmed bandit and the Gittins index*, J. Roy. Stat. Soc., 42 (1980), pp. 143-149.
[9] ———, *Optimization Over Time*, vol. 1, John Wiley, New York, 1982.
[10] E. B. DYNKIN, *Markov Processes*, vols. 1 and 2, Springer-Verlag, Berlin, New York, 1968.
[11] M. ROBIN, *Contrôle impulsionnel des processus de Markov*, Thèse, Paris IX, 1978.
[12] M. N. KAHETAKIS AND A. F. VEINOTT, *The multiarmed bandit problem: decomposition and computation*, Math. Oper. Res., 12 (1987), pp. 262-268.
[13] A. MANDELBAUM, *Discrete multiarmed bandit and multi-parameter processes*, Probab. Theory Related Fields, 71 (1986), pp. 129-147.
[14] ———, *Continuous multiarmed bandits and multi-parameter processes*, Ann. Probab., 15 (1987), pp. 1527-1556.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 113-136, January 1990

006

THE AUGMENTED LAGRANGIAN METHOD FOR PARAMETER ESTIMATION IN ELLIPTIC SYSTEMS*

KAZUFUMI ITO† AND KARL KUNISCH‡

Abstract. In this paper a new technique for the estimation of parameters in elliptic partial differential equations is developed. It is a hybrid method combining the output-least-squares and the equation error method. The new method is realized by an augmented Lagrangian formulation, and convergence as well as rate of convergence proofs are provided. Technically the critical step is the verification of a coercivity estimate of an appropriately defined Lagrangian functional. To obtain this coercivity estimate a seminorm regularization technique is used.

Key words. augmented Lagrangian method, parameter estimation, least squares, elliptic system

AMS(MOS) subject classifications. 35R30, 49D29

1. Introduction. In this paper we consider the problem of determining the unknown functional coefficient q in the elliptic partial differential equation

$$-\operatorname{div}(q \operatorname{grad} u) = f \quad \text{in } \Omega \qquad u = 0 \quad \text{on } \Gamma, \tag{1.1}$$

from an observation z of the solution u, where Ω is a bounded domain in $\mathbb{R}^n$, $n = 1, 2$, or 3, with piecewise smooth boundary Γ and $f \in H^{-1}$ is given. In applications, the function z might be constructed by interpolation of pointwise measurements. We propose and analyze a hybrid method that combines the output least squares and the equation error formulation [2], [17] within the mathematical framework given by the augmented Lagrangian technique.

The output least squares (OLS) approach is used most commonly and in our example for $n = 2$ or 3, it is stated, for instance, as the minimization problem in H^2:

$$\text{Minimize} \quad \frac{1}{2}|u(q) - z|_H^2 + \frac{\beta}{2} N(q) \tag{1.2}$$

$$\text{over} \quad Q_{\text{ad}} = \{q \in H^2(\Omega) : q \geqq \alpha \text{ and } |q|_{H^2} \leqq \gamma\}$$

where α and γ are positive constants chosen a priori, $u(q)$ is the solution of (1.1), and H is chosen as H^i, $i = 0$ or 1, for example. The second term in the cost functional represents a regularization term and Q_{ad} is chosen so that (1.2) has a solution for every $\beta \geqq 0$. The use of a regularization term guarantees the continuity of the mappings from the observation $z \in H$ ($H = H_0^1$ or L^2, for example) to a solution $q^\beta(z) \in Q_{\text{ad}} \subset \mathrm{H}^2$ for an appropriate choice of N and $\beta > 0$ and, in general, β cannot be taken equal to zero [3], [4], [9], [17]. In this paper we will use a regularization term such that $(N(q))^{1/2}$ is a seminorm on H^2. The use of seminorm regularization is very common for the inversion of linear operators [6], but it is not well studied in nonlinear problems such

* Received by the editors December 21, 1987; accepted for publication (in revised form) January 25, 1989.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported in part by U.S. Air Force Office of Scientific Research grants AFOSR-84-0398 and AFOSR-85-0303, National Aeronautics and Space Administration grant NAG-1-1517, and National Science Foundation grant UINT-8521208.

‡ Institut für Mathematik, Technische Universität Graz, Graz, Austria. The research of this author was supported in part by the Fonds zur Förderung der wissenschaftlichen Forschung, under grants S3206 and P6005, and by the U.S. Air Force Office of Scientific Research grant AFOSR-84-0398. Part of this work was performed while the author was visiting the Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

as the one presented by estimating q in (1.1). The OLS formulation is quite flexible with regard to the availability of the data. The OLS term in (1.2) can be adjusted in case data are only available over a subset of the domain, or are given as point measurements or as measurements of the flux at the boundary. As its form indicates, the OLS approach is less sensitive with respect to noise in the data when compared to the equation error method to be specified below. However, the minimization in (1.2) is an indirect method to determine the unknown q, and any iterative algorithm for solving (1.2) requires the solution of (1.1) for every update of $q \in Q_{\text{ad}}$.

An alternative to the OLS formulation is the equation error formulation. For our problem it can be stated as follows:

$$\text{(1.3)} \qquad \text{Minimize} \quad \tfrac{1}{2}|\nabla \cdot (q\nabla z) + f|_H^2$$

$$\text{subject to} \quad q \in H^2, \quad q \geqq \alpha$$

where H is either $H^{-1}(=(H_0^1)^*)$ or H^0. Since in computations H^{-1} requires a lesser amount of numerical differentiations of z, it should be preferred. An obvious disadvantage of this formulation is that it needs a fairly accurate observation of z defined over the entire domain Ω and it may be sensitive to noise in the data. On the other hand, it leads to efficient algorithms, since the minimization in (1.3) is quadratic.

The hybrid method that we propose not only combines both these formulations, but it also inherits the flexibility of the OLS approach and the quadratic structure of the equation error approach. This is achieved by viewing (1.2) as the following constrained minimization problem:

$$\text{(1.4)} \qquad \text{minimize} \quad F(q, u) = \frac{1}{2}|u - z|_{H_0^1}^2 + \frac{\beta}{2} N(q)$$

subject to

$$\text{(1.5)} \qquad -\nabla \cdot (q\nabla u) = f \quad \text{in } H^{-1},$$

$$\text{(1.6)} \qquad |q|_{H^2} \leqq \gamma,$$

$$\text{(1.7)} \qquad \alpha \leqq q \quad \text{on } \Omega,$$

in the two independent variables q and u. To solve (1.4)–(1.7) we apply the augmented Lagrangian algorithm (see, e.g., [1], [7], [8], [16]). It essentially involves minimizing a sequence of functionals of the form

$$\text{(1.8)} \qquad L_{c_k}(q, u; \lambda^k) = F(q, u) + \langle \lambda^k, e(q, u) \rangle_{H_0^1} + \frac{c_k}{2}|e(q, u)|_{H_0^1}^2 \quad \text{over } q \in Q_{\text{ad}},$$

and the multiplier sequence $\{\lambda^k\}$ in H_0^1 is given by

$$\text{(1.9)} \qquad \lambda^{k+1} = \lambda^k + c_k e(q_k, u_k),$$

where $\Delta: H_0^1 \to H^{-1}$ is the Laplacian, the function $e: H^2 \times H_0^1 \to H_0^1$ is defined by

$$\text{(1.10)} \qquad e(q, u) = (-\Delta)^{-1}(\nabla \cdot (q\nabla u) + f),$$

and the pair (q_k, u_k) minimizes (1.8). To carry out this iterative scheme a (possibly constant) sequence of positive real numbers $\{c_k\}$ and a startup $\lambda^1 \in H_0^1$ for the Lagrange multiplier need to be chosen. We suggest $\lambda^1 = 0$ but convergence will be guaranteed for any other choice of λ^1 as well. The inequality constraint $|q|_{H^2} \leqq \gamma$ (see (1.6)) can be augmented in a manner similar to the equality constraint $e(q, u) = 0$ (see § 2 for details). Convergence of this algorithm will be shown in Theorem 2.2 by employing

the results of [8], where a general framework for the analysis of the augmented Lagrangian method in infinite-dimensional spaces with equality constraints, as well as inequality constraints with finite-dimensional image space, is given. More precisely, convergence and rate of convergence of the pair (q_k, u_k) to a solution (q^β, u^β) of (1.4)-(1.7) in $H^2 \times H_0^1$, as well as of λ^k to λ^*, the Lagrange multiplier associated with the equality constraint (1.5), in H_0^1 will be proved. This result will be obtained under the assumption that the H^1-error between the data z and the nonregularized OLS-solution u^0 is sufficiently small and that the penalty parameters c_k are sufficiently large. It is not required that $\lim c_k = \infty$ as $k \to \infty$.

A number of remarks are in order.

(1) The minimization of the function in (1.8) requires the solution of a Poisson equation. For the discretized problem several efficient numerical techniques are readily available and any variant can be chosen. As a comparison, in the OLS approach (1.1) must be solved for $u = u(q)$ whenever a change in q occurs.

(2) Note that

$$|\varphi|^2_{H^{-1}} = \langle (-\Delta)^{-1}\varphi, \varphi\rangle = |(-\Delta)^{-1}\varphi|^2_{H_0^1} \quad \text{for all } \varphi \in H^{-1} = (H_0^1)^*.$$

Thus the minimization of the cost functional in (1.8) is a combination of the OLS-problem (1.2) and the equation error problem (1.3) where $H = H^{-1}$, with the aid of the multiplier method. The choice of the H_0^1-topology for the OLS-term and the H^{-1}-topology for the equation error term (equality constraint (1.5)) is natural from the point of view of the second-order sufficient optimality condition for (1.4) that will be used below, and the choice of these topologies leads to a method that requires the same amount of numerical differentiations in both the OLS and the equation error term.

(3) Note that $e(q, u)$ is a bilinear function in q and u. Thus for fixed q (respectively, u) (1.8) becomes quadratic in u (respectively, q) and we can take advantage of this structure numerically (see § 4 for details).

(4) As will be shown the penalty term $(c/2)|e(q, u)|^2_{H_0^1}$ enhances the convexity in the neighborhood of local minima.

(5) There is some arbitrariness in the choice of the topologies for the OLS and the equation error term. We can use $|\nabla(q\nabla u)+f|^2$ instead of $|e(q, u)|^2_{H_0^1}$ while simultaneously $|u-z|^2_{H_0^1}$ is replaced by $|u-z|^2_{H_0^1 \cap H^2}$. This would require a different analysis of the coercivity estimate (see § 3) and would lead to a different numerical implementation. This aspect is not exploited further within this paper.

The paper is organized as follows. In § 2 we describe in detail the augmented Lagrangian algorithm for the estimation of q in (1.1) and state the convergence results in Theorems 2.2 and 2.3. The essential technical tool that guarantees convergence is the positivity of the second Fréchet-derivative of the Lagrange functional (see (2.4) below). This coercivity condition is analyzed for various situations (Propositions 3.4 and 3.5) in § 3. Section 4 is devoted to a brief summary of the numerical experience that has been obtained with the augmented Lagrangian algorithm in parameter estimation. A comprehensive study of our numerical experience will appear elsewhere.

The notation that we use is rather standard. Unless otherwise specified, all function spaces are considered over the domain Ω. We use $\langle \cdot, \cdot \rangle$ to denote the inner product in L^2 and $|\cdot|$ to denote the norm in L^2 and $\mathbb{R}^n$, $n \geqq 1$. For other inner products and norms we use an index, as for instance $|v|_{H^1}$ denotes the common norm in H^1. The space $\mathbb{R}^n$ is endowed with the Euclidean norm. The inner product in H_0^1 is given by $\langle v, w\rangle_{H_0^1} = \langle \nabla v, \nabla w\rangle$ and the associated norm is defined through $|v|^2_{H_0^1} = \langle v, v\rangle_{H_0^1}$.

2. Problem formulation and convergence results. Let us formulate the problem for the multidimensional case ($n = 2$ or 3) first. We consider the constrained minimization

problem in the variables $(q, u) \in H^2 \times H_0^1$:

$$(P^\beta) \qquad \text{Minimize} \quad \frac{1}{2}|u-z|^2_{H_0^1} + \frac{\beta}{2}\left(\sum_{i_1,i_2} |q_{x_{i_1}x_{i_2}}|^2 + |\nabla q|^2\right),$$

$$\text{subject to} \quad e(q,u) = (-\Delta)^{-1}(\nabla \cdot (q\nabla u) + f) = 0 \text{ in } H_0^1, \text{ and}$$

$$|q|_{H^2} \leqq \gamma, q \geqq \alpha \quad \text{on } \Omega$$

where Δ is the Laplacian (considered as operator from H_0^1 onto H^{-1}), $f \in H^{-1}$, $z \in H_0^1$, $\beta \geqq 0$, and α, γ are given constants satisfying $\alpha^2 \int_\Omega dx < \gamma^2$. Note that (q^β, u^β) is a global solution of (P^β) if and only if q^β is a solution of (1.2) with $u(q^\beta) = u^\beta$. To argue existence of a solution q^β of (1.2) observe that Q_{ad} is a bounded, closed, and convex subset of H^2 and hence it is weakly sequentially compact. The parameter-to-solution mapping $q \to u(q)$, $q \in Q_{ad}$, is continuous from the space of continuous functions C to the H_0^1-topology. Moreover, H^2 is compactly embedded in C and norms are weakly lower semicontinuous functionals. Hence it can be shown that for any $\beta \geqq 0$ there exists at least one solution to (1.2) or equivalently (P^β).

Subsequently we will use the closed convex cone $\mathscr{C}$ with vertex at the origin of H^2 defined by

$$\mathscr{C} = \{w \in H^2 : w \leqq 0\}.$$

Let

$$\mathscr{C}^+ = \{\phi \in H^2 : \langle \phi, h \rangle_{H^2} \leqq 0 \text{ for all } h \in \mathscr{C}\}$$

be the positive dual cone and let

$$\mathbb{R}^- = \{x \in \mathbb{R} : x \leqq 0\}.$$

Then (P^β) can be written as follows:

$$(2.1) \qquad \text{Minimize} \quad F(q,u) = \frac{1}{2}|u-z|^2_{H_0^1} + \frac{\beta}{2} N(q)$$

$$\text{subject to} \quad e(q,u) = (-\Delta)^{-1}(\nabla \cdot (q\nabla u) + f) = 0,$$

$$g(q) = \tfrac{1}{2}(|q|^2_{H^2} - \gamma^2) \in \mathbb{R}^-,$$

$$l(q) = \alpha - q \in \mathscr{C}.$$

Henceforth (q^β, u^β) denotes a solution of (2.1). The next theorem (that will be proved in the latter part of this section) shows the existence and the uniqueness of a Lagrange multiplier $(\lambda^*, \mu^*, \eta^*) \in H_0^1 \times R^+ \times \mathscr{C}^+$ associated with a solution (q^β, u^β) of (P^β). We suppress the dependence of $(\lambda^*, \mu^*, \eta^*)$ on β.

THEOREM 2.1. *There exists a Lagrange multiplier* $(\lambda^*, \mu^*, \eta^*) \in H_0^1 \times \mathbb{R}^+ \times \mathscr{C}^+$ *such that*

$$L(q,u;\lambda^*,\mu^*,\eta^*) = F(q,u) + \langle \lambda^*, e(q,u) \rangle_{H_0^1} + \frac{\mu^*}{2}(|q|^2_{H^2} - \gamma^2) + \langle \eta^*, \alpha - q \rangle_{H^2}$$

satisfies

$$(2.2a) \qquad \nabla L(q^\beta, u^\beta; \lambda^*, \mu^*, \eta^*)(h,v) = 0 \quad \text{for all } (h,v) \in H^2 \times H_0^1$$

$$(2.2b) \qquad \mu^*(|q^\beta|^2_{H^2} - \gamma^2) = 0, \qquad \langle \eta^*, \alpha - q^\beta \rangle_{H^2} = 0.$$

Moreover, the Lagrange multiplier is unique and $\lambda^* = \Delta(q^\beta)^{-1}\Delta(u^\beta - z)$ *in* H_0^1 *where* $\Delta(q)u = \nabla \cdot (q\nabla u)$. *Here* $\nabla L(q^\beta, u^\beta; \lambda^*, \mu^*, \eta^*)(h,v)$ *denotes the Fréchet derivative of* $L(\cdot,\cdot;\lambda^*,\mu^*,\eta^*)$ *at* (q^β, u^β) *in direction* $(h,v) \in H^2 \times H_0^1$.

To solve (P^β) (or equivalently (2.1)) we will apply the augmented Lagrangian method. This method, due to Hestenes and Powell, has been studied extensively in the finite-dimensional case and, with equality constraints only, also in the infinite-dimensional case (cf. [1], [5], [7], and [16], for example). The infinite-dimensional case with equality as well as inequality constraints has been studied in [8]. To explain this method we require several reformulations of (P^β). For $c \geqq 0$ consider the augmented problem:

$$
\begin{aligned}
(P)_c \qquad \text{Minimize} \quad & F(q,u)+\frac{c}{2}|e(q,u)|^2_{H_0^1}+\frac{c}{2}|g(q)+w|^2 \\
\text{subject to} \quad & e(q,u)=0 \quad \text{in } H_0^1, \\
& g(q)+w=0, \qquad w\in\mathbb{R}, \\
& w \geqq 0, \\
& \alpha - q \in \mathscr{C}.
\end{aligned}
$$

In the notation of $(P)_c$ as well as F we suppress the dependence on β. We observe that (q^β, u^β) is a solution of (P^β) if and only if $(q^\beta, u^\beta, w^\beta = -g(q^\beta))$ is a solution of $(P)_c$. Moreover, it is simple to verify that $(\lambda^*, \mu^*, \mu^*, \eta^*)$ is a Lagrange multiplier for $(P)_c$, i.e.,

$$
\nabla L_c(q^\beta, u^\beta, w^\beta; \lambda^*, \mu^*, \eta^*)=0 \quad \text{and} \quad \langle \eta^*, \alpha - q^\beta\rangle_{H^2}=0, \quad \mu^* w^\beta = 0
$$

where the Lagrangian $L_c(q, u, w; \lambda^*, \mu^*, \eta^*)$ is given by

$$
\begin{aligned}
L_c(q,u,w;\lambda^*,\mu^*,\eta^*) &= F(q,u)+\langle \lambda^*, e(q,u)\rangle_{H_0^1}+\mu^* g(q) \\
&\quad +\langle \eta^*, \alpha - q\rangle_{H^2}+\frac{c}{2}|e(q,u)|^2_{H_0^1}+\frac{c}{2}|g(q)+w|^2,
\end{aligned} \tag{2.3}
$$

and f, e, g are defined in (2.1). It can also be shown that any solution $(q^\beta, u^\beta, w^\beta)$ of $(P)_c$ is a regular point in the sense of ([14], cf. also Theorem 2.1 and its proof), but since we will not use this fact we do not give its proof here.

Henceforth the following second-order sufficient optimality condition will be used:

There exist constants $\sigma > 0$ and $c_0 \geqq 0$ such that the second Fréchet derivative $\nabla^2 L_{c_0}$ of L_{c_0} satisfies the coercivity condition

$$
\nabla^2 L_{c_0}(q^\beta, u^\beta, w^\beta; \lambda^*, \mu^*, \eta^*)((h,v,y)(h,v,y)) \geqq \sigma(|h|^2_{H^2}+|v|^2_{H_0^1}+|y|^2) \tag{2.4}
$$

for all $(h, v, y) \in H^2 \times H_0^1 \times \mathbb{R}$, where the Lagrangian L_c is defined by (2.3) and $w^\beta = -g(q^\beta)$.

In the next section we will establish this condition for several specific cases. Under (2.4), it can be shown that $(q^\beta, u^\beta, w^\beta)$ is a solution of $(P)_c$ if and only if it is a solution of

$$
\min F(q,u)+\langle \lambda^*, e(q,u)\rangle_{H_0^1}+\mu^*(g(q)+w)+\frac{c}{2}|e(q,u)|^2_{H_0^1}+\frac{c}{2}|g(q)+w|^2, \tag{2.5}
$$

with $w \geqq 0$, $\alpha \leqq q$. In (2.5) the equality constraint and the inequality constraint with finite-dimensional image space are eliminated from the explicit constraints. However, (2.5) contains the unknown Lagrange multipliers (λ^*, μ^*). The augmented Lagrangian

algorithm applied to (P^β) involves solving iteratively for (λ^*, μ^*) and (q^β, u^β) and requires the solution of the following minimization problem:

(2.6) Given $\lambda \in H_0^1$ and $\mu \in \mathbb{R}^+$

$$\text{minimize} \quad F(q,u)+\langle \lambda, e\rangle_{H_0^1}+\mu(g(q)+w)+\frac{c}{2}|e(q,u)|^2_{H_0^1}+\frac{c}{2}|g(q)+w|^2$$

$$\text{subject to} \quad w \geqq 0,\ \alpha \leqq q \text{ and } (q,u,w) \in H^2 \times H_0^1 \times \mathbb{R}.$$

This problem is equivalent to the problem of minimizing

(2.7) $$F(q,u)+\langle \lambda, e\rangle_{H_0^1}+\mu\hat{g}(q,\mu,c)+\frac{c}{2}|e|^2_{H_0^1}+\frac{c}{2}\hat{g}(q,\mu,c)^2$$

$$= F(q,u)+\langle \lambda, e\rangle_{H_0^1}+\frac{c}{2}|e(q,u)|^2_{H_0^1}+\frac{1}{2c}(|\max(0, cg+\mu)|^2-|\mu|^2)$$

subject to $\alpha \leqq q$, where the constraint $w \geqq 0$ is eliminated. Here we put

$$\hat{g}(q,\mu,c)=\max\left(-\frac{\mu}{c}, g(q)\right)$$

and we used the equality

$$c\hat{g}(q,\mu,c)+\mu=\max(-\mu, cg(q))+\mu=\max(0, cg(q)+\mu).$$

Observe that $(q^\beta, u^\beta, w^\beta=\max(0, -g(q^\beta)-\mu/c))$ is a solution of (2.6) if and only if (q^β, u^β) is a solution of (2.7).

We are now prepared to specify the augmented Lagrangian algorithm to solve $(P)^\beta$. Choose a monotonically increasing sequence $\{c_k\}$ of positive real numbers with $c_1 > c_0$ and $(\lambda^1, \mu^1) \in H_0^1 \times \mathbb{R}^+$. In practice we suggest choosing $(\lambda^1, \mu^1)=(0,0)$, where the choice of λ^1 is based on Theorem 2.1: λ^* is close to 0 if u^β is close to z.

For $k \geqq 1$ determine (q_k, u_k) by solving the following:

(2.8) Minimize $\mathscr{L}_k(q,u)$,

$$\text{subject to} \quad (q,u) \in H^2 \times H_0^1,\ \alpha \leqq q,$$

and define

(2.9) $$\lambda^{k+1}=\lambda^k+(c_k-c_0)e(q_k,u_k), \qquad \mu^{k+1}=\mu^k+(c_k-c_0)\hat{g}(q_k,\mu^k,c_k)$$

where

$$\mathscr{L}_k(q,u)=F(q,u)+\langle \lambda^k, e\rangle_{H_0^1}+\frac{c_k}{2}|e|^2_{H^1}+\frac{1}{2c_k}(|\max(0, c_k g(q)+\mu^k)|^2-|\mu^k|^2).$$

In the following result we will ascertain local convexity of the cost functional appearing in (2.7) and (2.8) in a closed ball $\bar{B}$ containing the solution (q^β, u^β) of (P^β). We call (q_k, u_k) a solution of (2.8) in $\bar{B}$ if $\mathscr{L}_k(q_k,u_k) \leqq \mathscr{L}_k(q,u)$ for all $(q,u) \in \bar{B}$ with $\alpha \leqq q$. Existence of a solution of (2.8) in $\bar{B}$ and a Lagrange multiplier η^k associated with the inequality constraint $\alpha \leqq q$ can easily be verified. We will prove convergence of the solutions $(q_k, u_k) \in \bar{B}$ of (2.8) as $k \to \infty$.

It is useful to observe that

$$\mu^{k+1}=\max\left(\frac{\mu^k c_0}{c_k}, \mu^k+(c_k-c_0)g(q_k)\right)$$

and, since $\mu^1 \geqq 0$, this implies that $\mu^k \geqq 0$ for all $k \geqq 1$.

THEOREM 2.2. (a) *Suppose that the coercivity condition* (2.4) *holds. Then for every* $r \geqq \mu^*$ *there exist constants* $\tilde{c} = \tilde{c}(r) > c_0$ *and* $\bar{\sigma} > 0$, *and an open bounded ball* B *in* $H^2 \times H_0^1$ *centered at* (q^β, u^β) *such that*

$$F(q, u) + \langle \lambda^*, e \rangle_{H_0^1} + \mu^* \hat{g}(q, \mu, c) + \langle \eta^*, \alpha - q \rangle_{H^2} + \frac{c_0}{2} |e|_{H_0^1}^2 + \frac{c_0}{2} |\hat{g}(q, \mu, c)|^2 - F(q^\beta, u^\beta)$$

$$\geqq \bar{\sigma}(|q - q^\beta|_{H^2}^2 + |u - u^\beta|_{H_0^1}^2)$$

for all $(q, u) \in \bar{B}$, $c \geqq \tilde{c}$ *and* $\mu \in [0, r]$, *where* $\hat{g}(q, \mu, c) = \max(-\mu/c, g(q))$.

(b) *Suppose that in addition* $r \geqq \mu^* + (|\lambda^1 - \lambda^*|^2 + |\mu^1 - \mu^*|^2)^{1/2}$, *and* $c_1 \geqq \tilde{c}(r)$ *is chosen sufficiently large. Then every solution of* (2.8) *in* $\bar{B}$ *satisfies* $(q_k, u_k) \in B$, $\mu^k \in [0, r]$ *and*

$$\bar{\sigma}(|q_k - q^\beta|_{H^2}^2 + |u_k - u^\beta|_{H_0^1}^2) + \frac{1}{2(c_k - c_0)} (|\lambda^{k+1} - \lambda^*|_{H_0^1}^2 + |\mu^{k+1} - \mu^*|^2)$$

$$\leqq \frac{1}{2(c_k - c_0)} (|\lambda^k - \lambda^*|_{H_0^1}^2 + |\mu^k - \mu^*|^2), \tag{2.10}$$

for every $k \geqq 1$. *Moreover, there exists a constant* $K > 0$ *such that*

$$|q_k - q^\beta|_{H^2}^2 + |u_k - u^\beta|_{H_0^1}^2 \leqq \frac{1}{2\bar{\sigma}(c_k - c_0)} (|\lambda^k - \lambda^*|_{H_0^1}^2 + |\mu^k - \mu^*|^2) \quad \text{for } k \geqq 1,$$

$$|\lambda^k - \lambda^*|_{H_0^1}^2 + |\mu^k - \mu^*|^2 + |\eta^{k-1} - \eta^*|_{H^2}^2$$

$$\leqq \left(\frac{K}{\bar{\sigma}}\right)^{k-1} \prod_{i=1}^{k-1} \frac{1}{c_i - c_0} (|\lambda^1 - \lambda^*|_{H_0^1}^2 + |\mu^1 - \mu^*|^2) \quad \text{for } k \geqq 2. \tag{2.11}$$

The proof of Theorem 2.2 will be given in the latter part of this section.

Next we formulate the problem for the one-dimensional case. Without loss of generality we can assume that $\Omega = (0, 1)$. We take $(q, u) \in H^1 \times H_0^1$ and the regularization term $N(q)$ is chosen as

$$N(q) = \int_0^1 |q_x|^2 \, dx.$$

Thus, for $\beta \geqq 0$, the analogue of (P^β) is defined as follows:

$$\text{Minimize} \quad \frac{1}{2} |u - z|_{H_0^1}^2 + \frac{\beta}{2} |q_x|^2 \tag{2.12}$$

$$\text{subject to} \quad e(q, u) = (-\Delta)^{-1}((qu_x)_x + f) = 0 \quad \text{in } H_0^1,$$

$$|q|_{H^1} \leqq \gamma \quad \text{and} \quad \alpha \leqq q \text{ on } [0, 1]$$

where $\Delta: H_0^1 \to H^{-1}$ is given by $\Delta u = u_{xx}$. The results corresponding to Theorem 2.1 hold with $q \in H^2$ replaced by $q \in H^1$. In particular, if $(q^\beta, u^\beta) \in H^1 \times H_0^1$ is a solution of (2.12), then there exists a unique Lagrange multiplier $(\lambda^*, \mu^*, \eta^*) \in H_0^1 \times R^+ \times \mathscr{C}^+$ such that the Lagrangian

$$L(q; u; \lambda^*, \mu^*, \eta^*) = \tfrac{1}{2}|u - z|_{H_0^1}^2 + \frac{\beta}{2} |q_x|^2 + \langle \lambda^*, e \rangle_{H_0^1} + \frac{\mu}{2} (|q|_{H_0^1}^2 - \gamma^2) + \langle \eta^*, \alpha - q \rangle_{H^1}$$

satisfies $\nabla L(q^\beta, u^\beta; \lambda^*, \mu^*, \eta^*) = 0$ and $\mu^*(|q^\beta|_{H_0^1}^2 - \gamma^2) = \langle \eta^*, \alpha - q^\beta \rangle_{H^1} = 0$, where $\langle \cdot, \cdot \rangle_{H^1}$ is the duality pairing on H^1, $\mathscr{C} = \{h \in H^1 : h \leqq 0\}$ and $\mathscr{C}^+ = \{\varphi \in H^1 : \langle \varphi, h \rangle_{H^1} \leqq 0$

for all $h \in \mathscr{C}\}$ is the positive dual cone of $\mathscr{C}$. The Lagrange multiplier associated with the equality constraint can be expressed as

$$\lambda^* = \Delta(q^\beta)^{-1}\Delta(u^\beta - z) \quad \text{in } H_0^1$$

where $\Delta(q^\beta): H_0^1 \to H^{-1}$ is given by $\Delta(q^\beta)u = (q^\beta u_x)_x$. The Lagrangian for the augmented problem (compare (2.3)) is given by

$$L_c(q, u, w; \lambda^*, \mu^*, \eta^*) = \tfrac{1}{2}|u - z|^2_{H_0^1} + \frac{\beta}{2}|q_x|^2 + \langle \lambda^*, e\rangle_{H_0^1}$$

$$+ \langle \eta^*, \alpha - q\rangle_{H^1} + \mu^* g(q) + \frac{c}{2}|e|^2_{H_0^1} + \frac{1}{2}|g(q) + w|^2 \tag{2.13}$$

where $e = e(q, u)$, and $g(q) = \frac{1}{2}(|q|^2_{H^1} - \gamma^2)$.

The augmented Lagrangian method for the solution of (2.12) now proceeds precisely as in the multidimensional case described in (2.8), (2.9) with $q \in H^2$ replaced by $q \in H^1$, and the regularization term is chosen as $\beta|q_x|^2$. In particular, $\hat{g}(q, \mu, c) = \max(-\mu/c, \frac{1}{2}(|q|^2_{H^1} - \gamma))$ and (2.8) becomes

$$\begin{aligned} &\text{Minimize} \quad \mathscr{L}_k(q, u) \\ &\text{subject to} \quad (q, u) \in H^1 \times H_0^1, \alpha \leqq q, \end{aligned} \tag{2.14}$$

where

$$\mathscr{L}_k(q, u) = \frac{1}{2}|u - z|^2_{H_0^1} + \frac{\beta}{2}|q_x|^2 + \langle \lambda^k, e\rangle_{H_0^1} + \frac{c}{2}|e|^2_{H_0^1}$$

$$+ \frac{1}{2c_k}(|\max(0, c_k g(q) + \mu^k)|^2 - |\mu^k|^2).$$

The following analogue of Theorem 2.2 holds for the one-dimensional case.

THEOREM 2.3. *Suppose that there exist constants $\sigma > 0$ and $c_0 \geqq 0$ such that the second Fréchet derivative $\nabla^2 L_{c_0}$ of L_{c_0} satisfies*

$$\nabla^2 L_{c_0}(q^\beta, u^\beta, w^\beta; \lambda^*, \mu^*, \eta^*)((h, v, y), (h, v, y)) \geqq \sigma(|h|^2_{H^1} + |v|^2_{H_0^1} + |y|^2) \tag{2.15}$$

for all $(h, v, y) \in H^1 \times H_0^1 \times \mathbb{R}$. Then for every $r \geqq \mu^*$ there exist constants $\tilde{c} = \tilde{c}(r) > c_0$ and $\bar{\sigma} > 0$, and an open ball B in $H^1 \times H_0^1$ centered at (q^β, u^β) such that

$$\frac{1}{2}|u - z|^2_{H_0^1} + \frac{\beta}{2}|q_x|^2 + \langle \lambda^*, e\rangle_{H_0^1} + \mu^* \hat{g}(q, \mu, c) + \langle \eta^*, \alpha - q\rangle_{H^1} + \frac{c_0}{2}|e|^2_{H_0^1} + \frac{c_0}{2}|\hat{g}(q, \mu, c)|^2$$

$$\geqq \frac{1}{2}|u^\beta - z|^2_{H_0^1} + \frac{\beta}{2}|q_x^\beta|^2 + \bar{\sigma}(|q - q^\beta|^2_{H^1} + |u - u^\beta|^2_{H_0^1})$$

for all $(q, u) \in \bar{B}$, $c \geqq \tilde{c}$, and $\mu \in [0, r]$. Similarly, the assertions analogous to Theorem 2.2(b) *hold with H^2 replaced by H^1.*

We now come to the proofs of the results of this section.

Proof of Theorem 2.1. Let $\tilde{M}: H^2 \times H_0^1 \to H_0^1 \times \mathbb{R} \times H^2$ be defined by

$$\tilde{M}(h, v) = ((-\Delta)^{-1}(\nabla \cdot (q^\beta \nabla v + h \nabla u^\beta)), \langle q^\beta, h\rangle_{H^2}, -h).$$

The Fréchet derivatives ∇e, ∇g, and ∇l at the minimizer (q^β, u^β) in direction (h, v) are given by

$$\nabla e(q^\beta, u^\beta)(h, v) = (-\Delta)^{-1}(\nabla \cdot (q^\beta \nabla v + h \nabla u^\beta)) \in H_0^1,$$

$$\nabla g(q^\beta, u^\beta)(h, v) = \langle q^\beta, h\rangle_{H^2} \in \mathbb{R}, \qquad \nabla l(q^\beta, u^\beta)(h, v) = -h \in H^2.$$

These are the coordinates of $\tilde{M}$. The existence of a Lagrange multiplier satisfying (2.2) will follow directly from the regular point condition [14, p. 100] that for the problem under consideration is given by

$$\begin{aligned}&\{\tilde{M}(h, v)+(0, r, k)+\lambda(0, \tfrac{1}{2}(|q^\beta|_{H^2}^2-\gamma^2), \alpha-q^\beta):\\ &\qquad (h, v)\in H^2\times H_0^1, r\in\mathbb{R}^+, -k\in\mathscr{C}, \lambda\in\mathbb{R}\}=H_0^1\times\mathbb{R}\times H^2.\end{aligned}\tag{2.16}$$

We thus turn to the verification of (2.16) and choose $(w_1, w_2, w_3)\in H_0^1\times\mathbb{R}\times H^2$ arbitrarily.

Let

$$k=w_3-\min w_3$$

where the minimum is taken over $\bar{\Omega}$ and is well defined since H^2 embeds continuously into C. Observe that $-k\in\mathscr{C}$. Further define

$$h=\lambda(\alpha-q^\beta)-\min w_3,$$

with $\lambda\in\mathbb{R}$ to be fixed below. Clearly, $h\in H^2$ and

$$-h+k+\lambda(\alpha-q^\beta)=w_3,$$

and thus, independently of $\lambda\in\mathbb{R}$, the third coordinate in (2.16) is satisfied for this choice of k and h. Next we consider the second coordinate of (2.16):

$$\langle q^\beta, h\rangle_{H^2}+r+\frac{\lambda}{2}(|q^\beta|_{H^2}^2-\gamma^2)=w_2$$

or by the choice of h

$$r+\lambda(\langle q^\beta, \alpha\rangle-\tfrac{1}{2}|q^\beta|_{H^2}^2-\tfrac{1}{2}\gamma^2)=w_2+\langle q^\beta, \min w_3\rangle.\tag{2.17}$$

Since $\alpha^2\int_\Omega dx<\gamma^2$, the factor $\langle q^\beta, \alpha\rangle-\frac{1}{2}|q^\beta|_{H^2}^2-\frac{1}{2}\gamma^2$ in (2.17) is negative and thus there exist $\lambda\in\mathbb{R}^+$ and $r\in\mathbb{R}^+$ that satisfy (2.17). With $(\lambda, h, -k, r)\in\mathbb{R}\in H^2\times\mathscr{C}\times\mathbb{R}^+$ fixed we turn to the first coordinate in (2.16) that requires solving

$$\nabla(q^\beta\nabla v)=-\Delta w_1-\nabla(h\nabla u^\beta)\quad\text{in } H^{-1},$$

for $v\in H_0^1$. This is clearly possible and hence (2.16) is verified.

Next we will show that

$$\lambda^*=\Delta(q^\beta)^{-1}\Delta(u^\beta-z)\quad\text{in } H_0^1 \text{ where } \Delta(q)u=\nabla(q\nabla u).\tag{2.18}$$

Note that

$$\begin{aligned}&\nabla L(q^\beta, u^\beta; \lambda^*, \mu^*, \eta^*)(h, v)\\ &=\langle\nabla(u^\beta-z), \nabla v\rangle+\beta\Big(\langle\nabla q^\beta, \nabla h\rangle+\sum_{i_1, i_2}\langle(q_k)_{x_{i_1}x_{i_2}}, h_{x_{i_1}x_{i_2}}\rangle\Big)\\ &\quad-\langle\nabla\lambda^*, h\nabla u^\beta+q^\beta\nabla v\rangle+\mu^*\langle q^\beta, h\rangle_{H^2}-\langle\eta^*, h\rangle_{H^2}=0\end{aligned}\tag{2.19}$$

for all $(h, v)\in H^2\times H_0^1$. Let $h=0$ and $v\in H_0^1$ be arbitrary. Then we have

$$\langle\Delta(u^\beta-z), v\rangle-\langle\nabla\cdot(q^\beta\nabla\lambda^*), v\rangle=0$$

for all $v\in H_0^1$. This implies (2.18) and the uniqueness of the Lagrange multiplier λ^*.

To show the uniqueness of Lagrange multipliers μ^* and η^*, assume that $(\lambda^*, \mu_i^*, \eta_i^*)$, $i=1, 2$ are two Lagrange multipliers satisfying (2.2a) and (2.2b). Let $\mu=\mu_1^*-\mu_2^*$ and $\eta=\eta_1^*-\eta_2^*$. By (2.19) and (2.2b) we find

$$\mu\langle q^\beta, h\rangle_{H^2}-\langle\eta, h\rangle_{H^2}=0\quad\text{for all } h\in H^2$$

and moreover

$$\langle \eta, \alpha - q^\beta \rangle_{H^2} = 0.$$

First consider the case of $q^\beta \equiv \alpha$. Since $\alpha^2 \int_\Omega dx < \gamma^2$, the norm constraint is not active and therefore $\mu_1^* = \mu_2^* = 0$ in this case. Moreover, by the first equation we find that $\eta = 0$ so that $\eta_1^* = \eta_2^*$. Next assume that q^β is not identically α. Putting $h = \alpha - q^\beta$ in the first equation and using the second equation, we obtain $\mu \langle q^\beta, \alpha - q^\beta \rangle_{H^2} = 0$ that implies $\mu = 0$. Thus, $\langle \eta, h \rangle_{H^2} = 0$ for all $h \in H^2$. This implies $\eta = 0$ and the proof is completed.

Remark 2.4. If $\mu^* = \beta = 0$ and $\eta^* = 0$, then

$$\langle \nabla(u^\beta - z), \nabla u^\beta \rangle = 0. \tag{2.20}$$

In fact, suppose that $\mu^* = \beta = 0$ and $\eta^* = 0$. Then (2.19) with $v = 0$ implies $\langle \nabla \lambda^*, h \nabla u^\beta \rangle = 0$ for all $h \in H^2$. This equation with $h = q^\beta$ and λ^* expressed as in (2.18) implies

$$\langle \Delta(q^\beta)^{-1} \Delta(u^\beta - z), \Delta(q^\beta) u^\beta \rangle = \langle \nabla(u^\beta - z), \nabla u^\beta \rangle = 0,$$

which is the desired equality.

To prove Theorem 2.2 we need the following lemma.

LEMMA 2.5. *The adjoint operator* $\tilde{M}^*: H_0^1 \times \mathbb{R} \times H^2 \to H^2 \times H_0^1$ *of* $\tilde{M}$ *is surjective. The kernel of* $\tilde{M}^*$ *is one-dimensional and it is characterized by*

$$\ker \tilde{M}^* = \operatorname{span} \{(0, 1, q^\beta)\}.$$

Moreover, $\tilde{M}\tilde{M}^*$ *has a bounded inverse as an operator on range* $\tilde{M}$.

Proof. It is simple to show that the range of $\tilde{M}$ is closed. Next suppose that $\tilde{M}(h, v) = 0$ for some $(h, v) \in H^2 \times H_0^1$. Then $h = 0$ and thus $(-\Delta)^{-1} \nabla \cdot (q^\beta \nabla v) = 0$. Since $(-\Delta)^{-1}: H^{-1} \to H_0^1$ is an isomorphism and since $q^\beta \geqq \alpha$, this implies that $v = 0$. Hence $\tilde{M}$ is injective and by the closed range theorem $\tilde{M}^*$ is surjective. A short calculation shows that the kernel of $\tilde{M}^*$ is given as the set of elements $(x, y, z) \in H^1 \times \mathbb{R} \times H^2$ that satisfy

$$\langle y q^\beta, h \rangle_{H^2} = \langle z, h \rangle_{H^2} \quad \text{for all } h \in H^2.$$

In particular, $\dim (\ker \tilde{M}^*) = 1$. By the closed range theorem, $(\ker \tilde{M}^*)^\perp = \text{range } \tilde{M}$ and

$$H^1 \times R \times H^2 = \ker \tilde{M}^* \oplus \text{range } \tilde{M}.$$

To show that $\tilde{M}\tilde{M}^*$ has a bounded inverse on range $\tilde{M}$, we observe that $\tilde{M}\tilde{M}^*$ is injective and surjective on range $\tilde{M}$. This completes the proof.

Now we turn to a proof of Theorem 2.2.

Proof of Theorem 2.2. The augmentability estimate in (a) follows from the proof of Theorem 2.1 and from Corollary 2.2 of [8] (with (2.4) replacing Theorem 2.1 of [8]). In addition, if the norm constraint is not active, then $\tilde{c}(r)$ and B can be chosen so that

$$g(q) + \frac{\mu}{c} < 0 \tag{2.21}$$

for all $\mu \leqq r$, $c \geqq \tilde{c}(r)$, and all $(q, u) \in \bar{B}$. Estimate (2.10) is a special case of Proposition 4.1 of [8]. We now turn to the proof of (2.11). First observe that by (2.19)

$$\langle \nabla(u^\beta - z), \nabla v \rangle + \beta \left(\langle \nabla q^\beta, \nabla h \rangle + \sum_{i_1, i_2} \langle q^\beta_{x_{i_1} x_{i_2}}, h_{x_{i_1} x_{i_2}} \rangle \right)$$
$$- \langle \nabla \lambda^*, h \nabla u^\beta + q^\beta \nabla v \rangle + \mu^* \langle q^\beta, h \rangle_{H^2} - \langle \eta^*, h \rangle_{H^2} = 0$$

for all $(h, v) \in H^2 \times H_0^1$ and the necessary optimality condition for (q_k, u_k) yields

$$\begin{aligned}(2.22)\qquad &\langle \nabla(u_k - z), \nabla v\rangle + \beta\left(\langle \nabla q_k, \nabla h\rangle + \sum_{i_1,i_2} \langle (q_k)_{x_{i_1}x_{i_2}}, h_{x_{i_1}x_{i_2}}\rangle\right) \\ &\quad - \langle \nabla \tilde{\lambda}^{k+1}, h\nabla u_k + q_k \nabla v\rangle + \tilde{\mu}^{k+1}\langle q^k, h\rangle_{H^2} - \langle \eta^k, h\rangle_{H^2} = 0\end{aligned}$$

for all $(h, v) \in H^2 \times H_0^1$, where

$$\tilde{\lambda}^{k+1} = \lambda^k + c_k e(q_k, u_k) \quad \text{and} \quad \tilde{\mu}^{k+1} = \mu^k + c_k \hat{g}(q_k, \mu^k, c_k).$$

Subtracting these two equalities, rearranging terms, and using the definition of $\tilde{M}$, we obtain

$$\begin{aligned}&\langle (\lambda^* - \tilde{\lambda}^{k+1}, \mu^* - \tilde{\mu}^{k+1}, \eta^* - \eta^k), \tilde{M}(h, v)\rangle_{H_0^1 \times \mathbb{R} \times H^2} \\ (2.23)\qquad &= \langle \nabla(u_k - u^\beta), \nabla v\rangle + \beta\left(\langle \nabla(q_k - q^\beta), \nabla h\rangle + \sum_{i_1,i_2} \langle (q_k)_{x_{i_1}x_{i_2}} - q^\beta_{x_{i_1}x_{i_2}}, h_{x_{i_1}x_{i_2}}\rangle\right) \\ &\quad - \langle \nabla \tilde{\lambda}^{k+1} \cdot \nabla(u^\beta - u_k), h\rangle + \langle (q^\beta - q_k)\nabla \tilde{\lambda}^{k+1}, \nabla v\rangle + \tilde{\mu}^{k+1}\langle q^k - q^\beta, h\rangle_{H^2}.\end{aligned}$$

From (2.10), the sequences $\{\lambda^k, \mu^k\}$ and $\{q_k, u_k\}$ are uniformly bounded in $H_0^1 \times \mathbb{R}$ and $H^2 \times H_0^1$, respectively. This implies uniform boundedness of the sequence $\{\tilde{\lambda}^k, \tilde{\mu}^k\}$ in $H_0^1 \times \mathbb{R}$. By the Riesz Representation Theorem, the right-hand side of (2.23) can be represented by $\langle b_k, (h, v)\rangle_{H^2 \times H_0^1}$ where $b_k \in H^2 \times H_0^1$ and

$$|b_k|_{H^2 \times H_0^1} \leqq K_1 |(q_k, u_k) - (q^\beta, u^\beta)|_{H^2 \times H_0^1}$$

for a constant K_1 independent of k. We find from (2.23) that

$$\tilde{M}^*(\lambda^* - \tilde{\lambda}^{k+1}, \mu^* - \tilde{\mu}^{k+1}, \eta^* - \eta^k) = b_k$$

in $H^2 \times H_0^1$. If P_R denotes the orthogonal projection of $H_0^1 \times R \times H^2$ onto $(\ker \tilde{M}^*)^\perp$, then

$$\tilde{M}\tilde{M}^* P_R(\lambda^* - \tilde{\lambda}^{k+1}, \mu^* - \tilde{\mu}^{k+1}, \eta^* - \eta^k) = \tilde{M} b_k.$$

Since from Lemma 2.5 $\tilde{M}\tilde{M}^*$ is continuously invertible on range $\tilde{M} = (\ker \tilde{M}^*)^\perp$,

$$(2.24)\qquad |P_R(\lambda^* - \tilde{\lambda}^{k+1}, \mu^* - \tilde{\mu}^{k+1}, \eta^* - \eta^k)|_{H_0^1 \times \mathbb{R} \times H^2} \leqq K_2 |(q_k, u_k) - (q^\beta, u^\beta)|_{H^2 \times H_0^1},$$

for a constant K_2 independent of k.

Next the complementarity conditions imply that $\langle \eta^*, \alpha - q^\beta\rangle_{H^2} = 0$ and $\langle \eta^k, \alpha - q_k\rangle_{H^2} = 0$. Since $\{\tilde{\lambda}^k, \tilde{\mu}^k\}$ is uniformly bounded in $H_0^1 \times \mathbb{R}$, it follows from (2.22) that $\{\eta^k\}$ is uniformly bounded. Thus, these equalities yield the estimate

$$(2.25)\qquad |\langle \eta^* - \eta^k, \alpha - q^\beta\rangle_{H^2}| = |\langle \eta^k, q^\beta - q_k\rangle_{H^2}| \leqq K_3 |q^\beta - q_k|_{H^2},$$

for a constant K_3 independent of k. Now let us assume that $q^\beta \not\equiv \alpha$. Then $\langle q^\beta, \alpha - q^\beta\rangle_{H^2} \neq 0$ and it follows from (2.24), (2.25), and the characterization of $\ker \tilde{M}^*$ in Lemma 2.5 that

$$(2.26)\qquad |(\lambda^* - \tilde{\lambda}^{k+1}, \mu^* - \tilde{\mu}^{k+1}, \eta^* - \eta^k)|_{H_0^1 \times \mathbb{R} \times H^1} \leqq K_4 |(q_k, u_k) - (q^\beta, u^\beta)|_{H^2 \times H_0^1},$$

for a constant K_4 independent of k. In the case $q^\beta \equiv \alpha$ the norm constraint is inactive and hence $\mu^* = 0$. By (2.10) we have $\mu^k \leqq r$ for all $k \geqq 1$, and (2.21) then implies that $\tilde{\mu}^k = 0$ for all $k \geqq 2$. Consequently, (2.26) also holds for the case $q^\beta \equiv \alpha$. Next we will show that (2.26) holds when $(\tilde{\lambda}^{k+1}, \tilde{\mu}^{k+1})$ is replaced by $(\lambda^{k+1}, \mu^{k+1})$. Observe that there exists a constant K independent of k such that

$$|\lambda^{k+1} - \tilde{\lambda}^{k+1}|_{H_0^1} \leqq c_0 |\Delta^{-1}\nabla \cdot (q_k \nabla u_k - q^\beta \nabla u^\beta)|_{H_0^1} \leqq K |(q_k, u_k) - (q^\beta, u^\beta)|_{H^2 \times H_0^1},$$

and

$$|\mu^{k+1}-\tilde{\mu}^{k+1}| = c_0\left|\max\left(-\frac{\mu^k}{c_k}, g(q_k)\right)\right| \leqq K|q_k - q^\beta|_{H^2}$$

where for the last estimate we assumed that the norm constraint is active and we used the fact that $\mu^k \geqq 0$. Combining these estimates with (2.26), we obtain

$$|(\lambda^* - \lambda^{k+1}, \mu^* - \mu^{k+1}, \eta^* - \eta^k)|_{H_0^1 \times \mathbb{R} \times H^2} \leqq K_5 |(q_k, u_k) - (q^\beta, u^\beta)|_{H^2 \times H_0^1} \tag{2.27}$$

for a constant K_5 independent of k, provided that $|q^\beta|_{H^2} = \gamma$. If the norm constraint is not active, then by (2.10) we have

$$\mu^{k+1} \leqq \mu^* + (|\lambda^1 - \lambda^*|^2_{H_0^1} + |\mu^1 - \mu^*|^2)^{1/2} \leqq r \quad \text{for all } k \geqq 1$$

and therefore by (2.21)

$$\begin{aligned} \mu^{k+1} &= \max\left(\frac{\mu^k c_0}{c_k}, \mu^k + (c_k - c_0) g(q_k)\right) \\ &\leqq \max\left(\frac{\mu^k c_0}{c_k}, \mu^k - (c_k - c_0)\frac{\mu^k}{c_k}\right) = \frac{\mu^k c_0}{c_k}. \end{aligned} \tag{2.28}$$

The estimates (2.11) now follow from (2.10), (2.27), and (2.28). This completes the proof.

Remark 2.6. The assumption of Theorem 2.2(b) that $c_1 \geqq \tilde{c}(r)$ is sufficiently large is used to guarantee that (q_k, u_k) is in the open ball B for all $k \geqq 1$. If we only assume $c_1 \geqq \tilde{c}(r)$, then the estimates of Theorem 2.2 need to be modified. First (2.10) holds with (q_k, u_k) replaced by $(\tilde{q}_k, \tilde{u}_k)$, where $(\tilde{q}_k, \tilde{u}_k)$ is a solution of (2.8) in $\bar{B}$ (compare [8]). In particular, this implies that $(\tilde{q}_k, \tilde{u}_k) \to (q^\beta, u^\beta)$ and that $\{(\lambda^k, \mu^k)\}_{k=1}^\infty$ is bounded in $H_0^1 \times R$. Let k_0 be chosen such that $(\tilde{q}_k, \tilde{u}_k) \in B$ for all $k \geqq k_0$. Then the analysis of Theorem 2.2 can be repeated to show that for $k \geqq 1$

$$\begin{aligned} &|\lambda^{k+k_0} - \lambda^*|^2_{H_0^1} + |\mu^{k+k_0} - \mu^*|^2 + |\eta^{k+k_0-1} - \eta^*|_{H^2} \\ &\quad \leqq \left(\frac{K}{\bar{\sigma}}\right)^k \prod_{i=k_0}^{k+k_0-1} \frac{1}{c_i - c_0} (|\lambda^{k_0} - \lambda^*|^2_{H_0^1} + |\mu^{k_0} - \mu^*|^2) \end{aligned}$$

The proof of Theorem 2.3 can be given along the lines of that for Theorem 2.2.

3. The coercivity condition. In this section we establish the coercivity condition (2.4) for specific cases. To achieve this goal it is necessary to study the behavior of the solutions (q^β, u^β) to (P^β) as $\beta \to 0^+$. Let

$$N(q) = \sum_{i_1, i_2} |q_{x_{i_1} x_{i_2}}|^2 + |\nabla q|^2, \tag{3.1}$$

representing the seminorm regularization.

LEMMA 3.1. *Let (q^β, u^β) be any solution of (P^β). For $\beta > \beta^0 \geqq 0$, we have*

$$|u^\beta - z|^2_{H_0^1} \leqq |u^{\beta_0} - z|^2_{H_0^1} + \beta(N(q^{\beta_0}) - N(q^\beta)), \tag{3.2}$$

$$\sup_{Q^\beta} N(q^\beta) \leqq \inf_{Q^{\beta_0}} N(q^{\beta_0}), \tag{3.3}$$

$$\sup_{U^{\beta_0}} |u^{\beta_0} - z|_{H_0^1} \leqq \inf_{U^\beta} |u^\beta - z|_{H_0^1} \tag{3.4}$$

where for $\beta \geqq 0$, $Q^\beta = \{q^\beta : (q^\beta, u^\beta)$ is a solution of $(P^\beta)\}$ and $U^\beta = \{u^\beta : (q^\beta, u^\beta)$ is a solution of $(P^\beta)\}$. If $\beta_n \to 0^+$ and $\{q^{\beta_n}\}$ is any sequence of corresponding solution of (P^{β_n}), then $\{q^{\beta_n}\}$ has a weak cluster point, and every weak cluster point is a solution of (P^0), and we have

$$\lim_{n \to \infty} \sup_{Q^{\beta_n}} N(q^{\beta_n}) = \min_{Q^0} N(q^0). \tag{3.5}$$

Moreover, every weak cluster point of a sequence of solutions q^{β_n} is a strong cluster point in H^2 and it is a minimum norm solution of (1.2).

Proof. The proof of (3.2)–(3.5) is a simple consequence of the above remark on the equivalence between (1.2) and (P^β) and the results in § 2 of [4]. In fact, (3.2) follows from (2.2) in [4], (3.3) and (3.4) from Lemma 2.2, and (3.5) from Lemma 2.3 of [4]. Assumption (A2), requiring existence of a minimizer of (P^0) in [4], is guaranteed by the properties of Q_{ad}. The coercivity assumption for N in (A1) of [4] is replaced by the norm constraint. We will show the last statement of the lemma. If q^{β_n} converges weakly to q, then from (3.5) $N(q^{\beta_n}) \to N(q)$ where q is a minimum norm solution of (P^0). Since the embedding from H^2 into L^2 is compact, q^{β_n} converges strongly to q in L^2. Thus we obtain

$$N(q^{\beta_n}) + |q^{\beta_n}|_{L^2} \to N(q) + |q|_{L^2}^2.$$

Since $N(q) + |q|_{L^2}^2$ defines a norm that is equivalent to the common H^2-norm [15, p. 13], this implies $|q^{\beta_n}|_{H^2}^2 \to |q|_{H^2}^2$ so that $\{q^n\}$ converges strongly to q in H^2.

From (3.2), (3.3), (3.5), and observing that $|u^0 - z|_{H_0^1}^2$ is independent of $u^0 \in U^0$, we find the following corollary to Lemma 3.1.

COROLLARY 3.2. *There exists a real-valued monotonically increasing function $\rho(\beta)$ with $\lim \rho(\beta) \to 0$ as $\beta \to 0^+$ such that for $\beta \geqq 0$*

$$\begin{aligned} \sup_{U^\beta} |u^\beta - z|_{H_0^1}^2 &\leqq |u^0 - z|_{H_0^1}^2 + \beta \left(\min_{Q^0} N(q^0) - N(q^\beta)\right) \\ &\leqq |u^0 - z|_{H_0^1}^2 + \beta\rho(\beta). \end{aligned}$$

The second Fréchet derivative of L_c at $(q^\beta, u^\beta, w^\beta)$ in direction $(h, v, y) \in H^2 \times H_0^1 \times \mathbb{R}$ appearing in (2.4) is given by

$$\begin{aligned} &\nabla^2 L_c(q^\beta, u^\beta, w^\beta; \lambda^*, \mu^*, \eta^*)((h, v, y), (h, v, y)) \\ &\quad = |v|_{H_0^1}^2 + \beta N(h) - 2\langle \nabla\lambda^*, h\nabla v\rangle \\ &\qquad + \mu^* |h|_{H^2}^2 + c|(-\Delta)^{-1}\nabla \cdot (q^\beta \nabla v + h\nabla u^\beta)|_{H_0^1}^2 + c|\langle q^\beta, h\rangle_{H^2} + y|^2 \end{aligned} \tag{3.6}$$

where we used

$$|(-\Delta)^{-1}\nabla\varphi|_{H_0^1}^2 = \langle (-\Delta)^{-1}\nabla\varphi, \nabla\varphi\rangle = \langle \nabla\Delta^{-1}\nabla\varphi, \varphi\rangle = \langle P\varphi, \varphi\rangle_{L^2(\Omega, \mathbb{R}^n)} \tag{3.7}$$

for $\varphi \in H^1(\Omega; \mathbb{R}^n)$, which can be verified by Green's formula [15, p. 28]. Here the operator $P = \text{grad}\, \Delta^{-1}\, \text{div}$ defines an orthogonal projection in $L^2(\Omega, \mathbb{R}^n)$.

To prove the coercivity condition (2.4) in Proposition 3.4 we use some well-known estimates [15, pp. 18, 20, 72].

LEMMA 3.3. *There exist positive constants K_i, $i = 1, 2, 3, 4$, depending only on Ω such that*

(a) $|\varphi|_{L^\infty} \leqq K_1 |\varphi|_{H^2}$ *for all* $\varphi \in H^2$,

(b) $|\varphi| \leqq K_2 |\nabla\varphi|$ *for all* $\varphi \in H^1$ *with* $\int_\Omega \varphi\, dx = 0$,

(c) $|\varphi|_{H^2} \leqq K_3 (\sum_{i_1, i_2} |\varphi_{x_{i_1} x_{i_2}}|^2 + |\nabla\varphi|^2)^{1/2}$ *for all* $\varphi \in H^2$ *with* $\int_\Omega \varphi\, dx = 0$.

The proposition also involves the constants α and γ that define Q_{ad}, the function $\rho(\beta)$ from Corollary 3.2, and the constant ω that is defined as follows. For $f \in H^{-1}$ with $f \neq 0$ and $q \in Q_{\text{ad}}$ we have

$$|f|_{H^{-1}} = \sup_{v \in H_0^1} \frac{|\langle q\nabla u, \nabla v\rangle|}{|v|_{H_0^1}} \leqq |q|_{L^\infty} |\nabla u| \leqq K_1 \gamma |u|_{H_0^1}$$

and therefore

$$|u|_{H_0^1} \geqq \frac{|f|_{H^{-1}}}{\gamma K_1} =: \omega > 0. \tag{3.8}$$

PROPOSITION 3.4. *Let $f \in H^{-1}$ satisfy $f \neq 0$ and let $k = (1 + K_2^2 + 4K_1^2K_3^2)^{-1}$. Let (q^0, u^0) be a solution of (P^0) and suppose that for a constant $\beta_0 \in (0, 1)$*

$$|u^0 - z|_{H_0^1}^2 < \beta_0 \left[\frac{\alpha^2 k}{K_1^2} \left(1 - \frac{4k}{\omega} K_1^2 \gamma^2 \beta_0 \right) - \rho(\beta_0) \right]. \tag{3.9}$$

Then there exists a nontrivial compact interval $I = [\underline{\beta}, \beta_0] \subset (0, 1)$ and constants $c_0 > 0$, $\sigma_0 > 0$ such that

$$\nabla^2 L_c(q^\beta, u^\beta, w^\beta; \lambda^*, \mu^*, \eta^*)((h, v, y), (h, v, y)) \geqq \sigma_0(|h|_{H^2}^2 + |v|_{H_0^1}^2 + |y|^2)$$

for all $c \geqq c_0$, $(h, v, y) \in H^2 \times H_0^1 \times \mathbb{R}$ and any solution (q^β, u^β) of (P^β) with $\beta \in I$. Moreover, if $u^0 = z$, then such a constant β_0 always exists and I can be chosen as any compact subset of $(0, \beta_0]$.

We point out that in Proposition 3.4 the solutions (q^0, u^0) and (q^β, u^β) are assumed to be global. This is necessary since the proof requires the estimates of Lemma 3.1 and Corollary 3.2 that are given for (global) solutions. The assumption regarding existence of β_0 such that (3.9) holds represents a smallness condition of the error between z and the nonregularized OLS solution u^0. All quantities appearing on the right-hand side of (3.9) except for $\rho(\beta)$ in principle can be given explicitly.

Proof. Define a quadratic form on $H^2 \times H_0^1$ by

$$\begin{aligned} M_c(h, v) &= |v|_{H_0^1}^2 + \beta N(h) - 2\langle \nabla \lambda^*, h \nabla v \rangle \\ &\quad + c\langle (-\Delta)^{-1} \nabla \cdot (q^\beta \nabla v + h \nabla u^\beta), \nabla (q^\beta \nabla v + h \nabla u^\beta) \rangle \end{aligned} \tag{3.10}$$

where $c \geqq 0$ and the dependence of $M_c(h, v)$ on β is suppressed. For $(h, v, y) \in H^2 \times H_0^1 \times \mathbb{R}$ it follows that

$$\nabla^2 L_c((h, v, y), (h, v, y)) = M_c(h, v) + c(\langle q^\beta, h \rangle_{H^2} + y)^2 + \mu^* |h|_{H^2}^2. \tag{3.11}$$

We first concentrate on M_c. By Theorem 2.1 and Lemma 3.3(a)

$$\begin{aligned} \langle \nabla \lambda^*, h \nabla v \rangle &= -\langle \lambda^*, \nabla \cdot (h \nabla v) \rangle = \langle \nabla \Delta (q^\beta)^{-1} \Delta (u^\beta - z), h \nabla v \rangle \\ &\geqq -|\nabla \Delta (q^\beta)^{-1} \Delta (u^\beta - z)| |h \nabla v| \\ &\geqq -\frac{K_1}{\alpha} |\Delta (u^\beta - z)|_{H^{-1}} |h|_{H^2} |v|_{H_0^1} \\ &= -\frac{K_1}{\alpha} |u^\beta - z|_{H_0^1} |h|_{H^2} |v|_{H_0^1}. \end{aligned} \tag{3.12}$$

By (3.7) and Lemma 3.3 (suppressing the index β)

$$\begin{aligned} \langle (-\Delta)^{-1} \nabla \cdot (q \nabla v + h \nabla u), \nabla \cdot (q \nabla v + h \nabla u) \rangle &= \langle P(q \nabla v + h \nabla u), q \nabla v + h \nabla u \rangle \\ &\geqq \tfrac{1}{2} \langle P(h \nabla u), h \nabla u \rangle - \langle P(q \nabla v), q \nabla v \rangle \\ &\geqq \tfrac{1}{2} \langle P(h \nabla u), h \nabla u \rangle - |q|_{L^\infty}^2 |\nabla v|^2 \\ &\geqq \tfrac{1}{2} \langle P(h \nabla u), h \nabla u \rangle - K_1^2 \gamma^2 |v|_{H_0^1}^2. \end{aligned} \tag{3.13}$$

Here and in (3.14) below we suppress the superscript β. Each $h \in L^2$ can be uniquely decomposed as $h = h_1 + h_2$, where $h_1 = \int_\Omega h(x)\, dx$ and $h_2 \in \{h \in L^2 : \int_\Omega h\, dx = 0\}$. Observe that h_1 and h_2 are orthogonal in L^2 and if $h \in H^2$, then $h_2 \in H^2$. By definition of P and by Lemma 3.3(a) we find

$$\begin{aligned} \langle P(h \nabla u), h \nabla u \rangle &= \langle P(h_1 \nabla u), h_1 \nabla u \rangle + 2\langle P(h_1 \nabla u), h_2 \nabla u \rangle + \langle P(h_2 \nabla u), h_2 \nabla u \rangle \\ &\geqq |h_1|^2 |\nabla u|^2 - 2|h_1| |h_2|_{L^\infty} |\nabla u|^2 \\ &\geqq |h_1|^2 |\nabla u|^2 - 2K_1 |h_1| |h_2|_{H^2} |\nabla u|^2. \end{aligned} \tag{3.14}$$

Let us put $\ell := |u^\beta|^2_{H_0^1}$ and $c = \delta\beta$ with $\delta > 0$ to be chosen below. Then from (3.10), (3.12)–(3.14) we have

$$M_{\delta\beta}(h, v) \geqq (1 - \delta\beta K_1^2\gamma^2)|v|^2_{H_0^1} + \beta\left(N(h) + \frac{\delta\ell}{2}|h_1|^2 - K_1\delta\ell|h_1||h_2|_{H^2}\right)$$

$$- \frac{2K_1}{\alpha}|u^\beta - z|_{H_0^1}|h|_{H^2}|v|_{H_0^1}. \tag{3.15}$$

Next we will show that there exist constants k and δ such that

$$N(h) + \frac{\delta\ell}{2}|h_1|^2 - K_1\delta\ell|h_1||h_2|_{H^2} \geqq k(N(h) + |h|^2) \tag{3.16}$$

for all $h \in H^2$.

From Lemma 3.3 it follows that for any $0 < A < 1$ and $B > 0$

$$N(h) + \frac{\delta\ell}{2}|h_1|^2 - K_1\delta\ell|h_1||h_2|_{H^2}$$

$$\geqq (1 - A)N(h) + \frac{\delta\ell}{2}|h_1|^2 + \frac{A}{K_2^2}|h_2|^2 - \frac{(\delta\ell)^2}{4}B^2K_1^2|h_1|^2 - \frac{1}{B^2}|h_2|^2_{H^2}$$

$$\geqq \left(1 - A - \frac{K_3^2}{B^2}\right)N(h) + \left(\frac{\delta\ell}{2} - \frac{(\delta\ell)^2}{4}B^2K_1^2\right)|h_1|^2 + \frac{A}{K_2^2}|h_2|^2.$$

Since $|h|^2 = |h_1|^2 + |h_1|^2$, inequality (3.16) holds if there exist positive constants δ, k, B, and $A \in (0, 1)$ such that

$$1 - A - \frac{K_3^2}{B^2} \geqq k, \quad \frac{\delta\ell}{2} - \frac{(\delta\ell)^2}{4}B^2K_1^2 \geqq k \quad \text{and} \quad \frac{A}{K_2^2} \geqq k.$$

A calculation shows that these inequalities are satisfied if we take

$$k = (1 + K_2^2 + 4K_1^2K_3^2)^{-1}, \quad A = K_2^2k, \quad B^2 = (4kK_1^2)^{-1}, \quad \delta\ell = 4k.$$

This is the choice of k contained in the statement of the proposition. From (3.15) and (3.16) we have

$$M_c(h, c) \geqq k\beta(N(h) + |h|^2) + \left(1 - \frac{4k}{\ell}\beta K_1^2\gamma^2\right)|v|^2_{H_0^1}$$

$$- \frac{2K_1}{\alpha}|u^\beta - z|_{H_1^0}|h|_{H^2}|v|_{H_0^1}$$

where $c = 4k\beta/\ell$ and $\ell = \ell(\beta) = |u^\beta|^2_{H_0^1}$. By the choice of ω the last inequality implies

$$M_c(h, v) \geqq k\beta|h|^2_{H^2} + \left(1 - \frac{4k}{\omega}\beta K_1^2\gamma^2\right)|v|^2_{H_0^1} - \varepsilon k\beta|h|^2_{H^2} - \frac{K_1^2}{\varepsilon\alpha^2k\beta}|u^\beta - z|^2_{H_0^1}|v|^2_{H_0^1}$$

$$= (1 - \varepsilon)k\beta|h|^2_{H^2} + \left(1 - \frac{4k}{\omega}\beta K_1^2\gamma^2 - \frac{K_1^2}{\varepsilon\beta\alpha^2k}|u^\beta - z|^2_{H_0^1}\right)|v|^2_{H_0^1} \tag{3.17}$$

where $c = 4k\beta/\omega$ and $\varepsilon \in (0, 1)$ is arbitrary. By (3.9) and Corollary 3.2 we find

$$\sup_{U^{\beta_0}} |u^{\beta_0} - z|^2_{H_0^1} < \frac{\varepsilon_1\beta_0\alpha^2k}{K_1^2}\left(1 - \frac{4k}{\omega}K_1^2\gamma^2\beta_0\right)$$

for some $\varepsilon_1 \in (0, 1)$. Furthermore, by (3.4) of Lemma 3.1 there exist constants $\bar{\beta} \in (0, \beta_0)$ and $\eta > 0$ such that

$$\sup_{U^\beta} |u^\beta - z|^2_{H_0^1} \leqq \frac{\varepsilon_1 \beta \alpha^2 k}{K_1^2} \left(1 - \frac{4k}{\omega} K_1^2 \gamma^3 \beta\right) - \eta$$

for all $\beta \in I = [\bar{\beta}, \beta_0]$ and by (3.17) with $\varepsilon = \varepsilon_1$ this implies

$$M_c(h, v) \geqq (1 - \varepsilon_1) k\beta |h|^2_{H^2} + \eta |v|^2_{H_0^1} \quad \text{where } c = \frac{4k\beta}{\omega}$$

for all $\beta \in I$. From the definition of M_c and the last inequality, it follows that there exists a constant $\sigma > 0$ such that

$$M_c(h, v) \geqq \sigma\beta(|h|^2_{H^2} + |v|^2_{H_0^1}) \quad \text{for all } c \geqq \frac{4k\beta}{\omega} \text{ and } \beta \in I. \tag{3.18}$$

If $u^0 = z$ then there always exists $\beta_0 > 0$ such that (3.9) is satisfied, and using Corollary 3.2 we can verify (3.18) with I any compact subset of $(0, \beta_0]$.

Next note that

$$|\langle q^\beta, h\rangle_{H^2} + y|^2 \geqq \tfrac{1}{2}|y|^2 - |\langle q^\beta, h\rangle_{H^2}|^2 \geqq \tfrac{1}{2}|y|^2 - \gamma^2 |h|^2_{H^2}$$

and consequently

$$\begin{aligned} \sigma\beta |h|^2_{H^2} + c_2 |\langle q^\beta, h\rangle_{H^2} + y|^2 &\geqq (\sigma\beta - \gamma^2 c_2)|h|^2_{H^2} + \frac{1}{2} c_2 |y|^2 \\ &= \frac{\sigma\beta}{1 + 2\gamma^2} (|h|^2_{H^2} + |y|^2) \end{aligned} \tag{3.19}$$

where $c_2 = 2\sigma\beta/(1 + 2\gamma^2)$. Hence from (3.11), (3.18), and (3.19) we obtain

$$\nabla^2 L_c(q^\beta, u^\beta, w^\beta; \lambda^*, \mu^*, \eta^*)((h, v, y), (h, v, y)) \geqq \frac{\sigma\beta}{1 + 2\gamma^2} (|h|^2 + |v|^2_{H^1} + |y|^2)$$

for all $(h, v, y) \in H^2 \times H_0^1 \times \mathbb{R}$, $c \geqq \max(4k\beta/\omega, 2\sigma\beta/(1 + 2\gamma^2))$ and $\beta \in I$. This implies the claim.

In special cases the coercivity estimate of Proposition 3.4 can be obtained with $\beta = 0$.

PROPOSITION 3.5. *Let (q^0, u^0) be a local solution of (P_0).*

(a) *If $\mu^* > 0$ and $\operatorname{dist}^2 := |u^0 - z|^2_{H_0^1} < (\alpha / K_1)^2 \mu^*$, then there exist positive constants σ_1 and c_1 such that*

$$\nabla^2 L_c(q^0, u^0, w^0)((h, v, y), (h, v, y)) \geqq \sigma_1 (|h|^2_{H^2} + |v|^2_{H_0^1} + |y|^2)$$

for all $c \geqq c_1$ and $(h, v, y) \in H^2 \times H_0^1 \times \mathbb{R}$.

(b) *Let $\tilde{L}_c$ be given by (1.8) where the norm constraint is not augmented (i.e., $L_c = \tilde{L}_c + (c/2)|g(q) + w|^2$). If $\operatorname{dist} = 0$, then there exist positive constants σ_2 and c_2 such that*

$$\nabla^2 \tilde{L}_c(q^0, u^0)((h, v), (h, v)) \geqq \sigma_2 (|P(h \nabla u^0)|^2_{L^2} + |v|^2_{H_0^1}) \tag{3.20}$$

for all $c \geqq c_2$ and $(h, v) \in H^2 \times H_0^1$.

(c) *Let $\{\varphi_i\}_{i=1}^M$ be (curved) linear elements [19] or indicator functions such that $0 \leqq \varphi_i \leqq 1$ on Ω and let $V^M = \{q = \sum_{i=1}^M q_i \varphi_i : q_i \in \mathbb{R}\} \subset L^\infty$. Let $(P^0)_c^M$ be the problem of minimizing*

$$\frac{1}{2}|u - z|^2_{H_0^1} + \frac{c}{2}|e(q, u)|^2_{H_0^1}$$

subject to $e(q, u) = 0$ *in* H_0^1 *and* $q \in Q_{\mathrm{ad}} = \{q = \sum_{i=1}^M q_i \varphi_i \in V^M : q_i \geqq \alpha$ *and* $q^T W^M q < \gamma^2\}$, *where* $|(W^M)^{1/2}\alpha|^2 < \gamma^2$ *and* W^M *is a symmetric positive definite matrix on* $\mathbb{R}^M$. *Furthermore, assume that* $q \in Q_{\mathrm{ad}}$ *implies* $q(x) \geqq \alpha$ *on* Ω. *Then* $(P^0)_c^M$ *has a solution* $(q^0, u^0) \in V^M \times H_0^1$ *with an associated Lagrange multiplier* $(\lambda^*, \mu^*, \eta^*) \in H_0^1 \times \mathbb{R}_+ \times \mathbb{R}_+^M$ *such that if we define the augmented Lagrangian* (*compare* (2.13))

$$\hat{L}_c(q, u, w; \lambda^*, \mu^*, \eta^*) = \frac{1}{2}|u - z|_{H_0^1}^2 + \langle \lambda^*, e\rangle_{H_0^1} + \frac{\mu^*}{2}(q^T W^M q - \gamma^2) + \sum_{i=1}^M \eta_i^*(\alpha - q_i) + \frac{c}{2}|e|_{H_0^1}^2 + \frac{c}{2}\left(\frac{1}{2}(q^T W^M q - \gamma^2) + w\right)^2,$$

then $\nabla \hat{L}_c(q^0, u^0, w^0; \lambda^*, \mu^*, \eta^*)(h, v, y) = 0$ *for all* $(h, v, y) \in V^M \times H_0^1 \times \mathbb{R}$ *and* $\mu^*(q^{0T} W^M q^0 - \gamma^2) = \sum_{i=1}^M \eta_i^*(\alpha - q_i^0) = 0$. *Moreover, if* $h \to |P(h\nabla u^0)|$ *defines a norm on* V^M *with* $|P(h\nabla u^0)| \geqq b|h|_{L^\infty}$ *for some* $b > 0$ *and all* $h \in V^M$, *and if* $\mathrm{dist} < \alpha b^2 (2|q^0|_{L^\infty}^2 + b^2)^{-1}$, *then there exist positive constants* σ_3 *and* c_3 *such that*

$$\nabla^2 \hat{L}_c(q^0, u^0, w^0; \lambda^*, \mu^*, \eta^*)((h, v, y), (h, v, y)) \geqq \sigma_3(|P(h\nabla u^0)|^2 + |v|_{H_0^1}^2 + y^2)$$

for all $c \geqq c_3$ *and* $(h, v, y) \in V^M \times H_0^1 \times \mathbb{R}$.

We precede the proof with a brief discussion of this proposition. Part (a) presents the most desirable situation. In this case $\mu^* > 0$ takes over the role of the regularization parameter $\beta > 0$ of Proposition 3.4. In general, however, it is difficult to give conditions that guarantee $\mu^* > 0$ (see, however, [13] for the one-dimensional case). Part (b) is not directly applicable for the results of § 2, but it exhibits clearly the difficulties that are involved in obtaining a lower bound on the second derivative of the augmented Lagrangian: First the norm involved in (3.20) is only the L^2 rather than the H^2-norm; second, we obtain an estimate only in terms of $P(h\nabla u^0)$, where the kernel of P is the set of all divergence free vector fields. However, (3.20) also indicates how further assumptions can be made to obtain the desired coercivity estimate. To give an example, let us assume that q is known to be constant a priori, i.e., we take $q \in \{q \in \mathbb{R}: \alpha \leqq q \leqq \gamma(\int_\Omega dx)^{1/2}\}$. Then $h \in \mathbb{R}$ and $P(h\nabla u^0)$ becomes $h\nabla u^0$ and the desired coercivity estimate holds, with the H^2-norm replaced by the norm in $\mathbb{R}$, if $f \neq 0$. A less trivial case is considered in part (c) of the proposition. In the statement of (c) we did not distinguish between a function q and its coordinate expansion q in terms of φ_i. Moreover, we used α to also stand for $\mathrm{col}\,(\alpha, \cdots, \alpha) \in \mathbb{R}^M$ and we recall that $2w^0 = \gamma^2 - q^{0T} W^M q^0$.

Proof of Proposition 3.5. (a) First observe that Theorem 2.1 is applicable for local solutions (q^0, u^0) of (P^0). From (3.10) with $\beta = 0$, (3.12) and (3.13) we conclude that

$$M_c(h, v) + \mu^*|h|_{H^2}^2 = |v|_{H_0^1}^2 - 2\langle \nabla\lambda^*, h\nabla v\rangle + c|P(q^0\nabla v + h\nabla u^0)|^2 + \mu^*|h|_{H^2}^2$$

$$\geqq |v|_{H_0^1}^2 - \frac{2K_1}{\alpha}\,\mathrm{dist}\,|h|_{H^2}|v|_{H_0^1} + \mu^*|h|_{H^2}^2$$

$$\geqq |v|_{H_0^1}^2(1 - \varepsilon) + |h|_{H^2}^2\left(\mu^* - \frac{K_1^2\,\mathrm{dist}^2}{\varepsilon\alpha^2}\right)$$

for any $\varepsilon > 0$. The assumption on dist implies the existence of $\sigma > 0$ such that

$$M(h, v) + \mu^*|h|_{H^2}^2 \geqq \sigma(|h|_{H^2}^2 + |v|_{H_0^1}^2)$$

for all $(h, v) \in H^2 \times H_0^1$. The claim now follows with the same argument as at the end of Proposition 3.4.

(b) Since dist $= 0$, Theorem 2.1 implies that $\lambda^* = 0$, and further

$$\nabla^2 \tilde{L}_c(q^0, u^0)((h, v), (h, v)) = |v|^2_{H_0^1} + c|P(q^0 \nabla v + h \nabla u^0)|^2 + \mu^* |h|^2_{H^2}$$

$$\geqq |v|^2_{H_0^1}(1 - c|q^0|^2_{L^\infty}) + \frac{c}{2}|P(h \nabla u^0)|^2$$

$$\geqq |v|^2_{H_0^1}(1 - cK_1^2 \gamma^2) + \frac{c}{2}|P(h \nabla u^0)|^2.$$

This estimate implies the claim.

(c) The assumptions on α, q, and φ_i imply that Q_{ad} is nonempty and that $q(x) \geqq \alpha$ for every $q \in Q_{\text{ad}}$. It is simple to argue existence of a solution (q^0, u^0) of $(P^0)_c^M$. Moreover, the conclusions of Theorem 2.1 and Lemma 2.5 remain valid if h and $q^\beta \in H^2$ are replaced by h and $q^0 \in \mathbb{R}^M$, if $\mathbb{R}^M$ is endowed with the inner product $\langle h, W^M h \rangle$, and if $\mathscr{C} = \mathbb{R}^M_-$, $\ell(q) = \alpha - q \in \mathbb{R}^M$. In particular, there exists a Lagrange muitiplier $(\lambda^*, \mu^*, \eta^*)$ with the specified properties.

For $h = \sum_{i=1}^M h_i \varphi_i \in V^M$ and $v \in H_0^1$ we find

$$M_c(h, v) := |v|^2_{H_0^1} + 2\langle \nabla \lambda^*, h \nabla u \rangle + c|P(q^0 \nabla v + h \nabla u^0)|^2$$

$$\geqq |v|^2_{H_0^1} - \frac{2}{\alpha} \operatorname{dist} |h|_{L^\infty} |v|_{H_0^1} + \frac{c}{2}|P(h \nabla u^0)|^2 - c|P(q^0 \nabla v)|^2$$

$$\geqq (1 - c|q^0|^2_{L^\infty})|v|^2_{H_0^1} - \frac{2}{\alpha b} \operatorname{dist} |v|_{H_0^1} |P(h \nabla u^0)| + \frac{c}{2}|P(h \nabla u^0)|^2.$$

If $\tilde{c}_3 = 2(b^2 + 2|q^0|_{L^\infty})^{-1}$ then for $c \geqq c_3$ the following inequality holds:

$$M_c(h, v) \geqq \sigma |v|^2_{H_0^1} - \frac{2}{\alpha b} \operatorname{dist} |v|_{H_0^1} |P(h \nabla u^0)| + \frac{\sigma}{b^2}|P(h \nabla u^0)|^2$$

where $\sigma = b^2(b^2 + 2|q^0|^2_{L^\infty})^{-1}$. Thus, if dist $< \sigma\alpha$, then there exists a constant $\tilde{\sigma}_3 > 0$ such that

$$M_c(h, v) \geqq \tilde{\sigma}_3(|v|^2_{H_0^1} + |P(h \nabla u^0)|^2) \tag{3.21}$$

for all $c \geqq \tilde{c}_3$. Since $\nabla \hat{L}_c^2(q^0, u^0, w^0; \lambda^*, \mu^*, \eta^*)((h, v, y)(h, v, y)) = M_c(h, v) + \mu^* h^T W^M h + c(q^{0^T} W^M h + y)^2$, the final claim follows from (3.21) and an argument analogous to the one at the end of Proposition 3.4.

Next we consider the one-dimensional case where we have an explicit formula for the orthogonal projection P and explicit values for the estimates in Lemma 3.3. By D we denote differentiation and the domain Ω is $(0, 1)$.

LEMMA 3.6. *The operator $P = D\Delta^{-1}D$ is an orthogonal projection on L^2. Moreover,* ker P *is the set of all constant functions on* $(0, 1)$ *and* $|P\varphi|^2 = |\varphi|^2 - (\int_0^1 \varphi \, dx)^2$ *for* $\varphi \in L^2$.

Proof. The first part of the lemma is obvious. Next recall that $\{1, \sqrt{2} \cos \pi x, \sqrt{2} \cos 2\pi x, \cdots\}$ is a complete orthonormal system in L^2. For $\varphi \in H^1$ with $\varphi = a_0 + \sum_{k=1}^\infty a_k \sqrt{2} \cos k\pi x$ it is easy to see that $P\varphi = \sum_{k=1}^\infty a_k \sqrt{2} \cos k\pi x$ and $(I - P)\varphi = a_0$. Since H^1 is dense in L^2 this follows for all $\varphi \in L^2$ and, moreover, $|P\varphi|^2 = |\varphi|^2 - |(I - P)\varphi|^2 = |\varphi|^2 - (\int_0^1 \varphi \, dx)^2$.

Lemma 3.7. (a) $|\varphi|_{L^\infty} \leqq \sqrt{2}|\varphi|_{H^1}$ *for all* $\varphi \in H^1$,

(b) $|\varphi|_{L^\infty} \leqq 1/\sqrt{3}|D\varphi|$ *for all* $\varphi \in H^1$ *with* $\int_0^1 \varphi \, dx = 0$,

(c) $\pi|\varphi| \leqq |D\varphi|$ *for all* $\varphi \in H^1$ *with* $\int_0^1 \varphi \, dx = 0$.

Proof. (a) Let $\varphi \in H^1$. By the Mean Value Theorem there exists $\zeta \in [0, 1]$ such that $\varphi(\zeta) = \int_0^1 \varphi(x) \, dx$. For every $x \in [0, 1]$ we have

$$\varphi(x) = \int_\zeta^x D\varphi(s) \, ds + \varphi(\zeta).$$

This implies

$$|\varphi(x)| \leqq \int_0^1 (|D\varphi(s)| + |\varphi(s)|)\, ds \leqq \left(\int_0^1 (|D\varphi(s) + |\varphi(s)|)^2\, ds \right)^{1/2} \leqq \sqrt{2}|\varphi|_{H^1}.$$

(b) By assumption, φ can be expressed as $\varphi = \sum_{k=1}^{\infty} a_k \sqrt{2} \cos k\pi x$. Therefore $|D\varphi|^2 = \pi^2 \sum_{k=1}^{\infty} k^2 a_k^2$, and

$$|\varphi|_{L^\infty} \leqq \sqrt{2} \sum_{k=1}^{\infty} |a_k| = \sqrt{2} \left(\sum_{k=1}^{\infty} k^2 |a_k|^2 \right)^{1/2} (\sum k^{-2})^{1/2}$$

$$= \sqrt{2} \frac{|D\varphi|}{\pi} \cdot \frac{\pi}{\sqrt{6}} = \frac{|D\varphi|}{\sqrt{3}},$$

which was to be proved.

(c) The assertion immediately follows from

$$|\varphi|^2 = \sum_{k=1}^{\infty} |a_k|^2 \quad \text{and} \quad |D\varphi|^2 = \pi^2 \sum_{k=1}^{\infty} k^2 |a_k|^2 \geqq \pi^2 \sum_{k=1}^{\infty} |\alpha_k|^2.$$

Analogously to the multidimensional case we find a lower bound on the solutions of $-D(qDu) = f$ for $q \in Q_{\text{ad}}$:

$$|f|_{H^{-1}} = \sup_{V \in H_0^1} \frac{\langle qDu, Dv \rangle}{|v|_{H_0^1}} \leqq |q|_{L^\infty} |u|_{H_0^1} \leqq \sqrt{2}|q|_{H^1} |u|_{H_0^1} \leqq \sqrt{2}\gamma |u|_{H_0^1},$$

and thus

$$|u(q)|_{H_0^1} \geqq \omega_1 := \frac{|f|_{H^{-1}}}{\sqrt{2}\gamma} \quad \text{for all } q \in Q_{\text{ad}}. \tag{3.22}$$

PROPOSITION 3.8. *Let $f \in H^{-1}$ satisfy $f \neq 0$ and let $k = 3\pi^2/(7\pi^2 + 3)$. Let (q^0, u^0) be a solution of (P^0) and suppose that for a constant $\beta_0 \in (0, 1)$*

$$|u^0 - z|_{H_0^1}^2 < \beta_0 \left[\frac{\alpha^2 k}{2} \left(1 - \frac{8k}{\omega_1} \gamma^2 B_0 \right) - \rho(\beta_0) \right]. \tag{3.23}$$

Then there exists a nontrivial compact interval $I = [\underline{\beta}, \beta_0] \subset (0, 1)$ and constants $c_0 > 0$, $\sigma_0 > 0$ such that

$$\nabla^2 L_c(q^\beta, u^\beta, w^\beta; \lambda^*, \mu^*, \eta^*)((h, v, y), (h, v, y)) \geqq \sigma_0 (|h|_{H^1}^2 + |v|_{H_0^1}^2 + |y|^2)$$

for all $c \geqq c_0$, $(h, v, y) \in H^2 \times H_0^1 \times \mathbb{R}$ and any solution (q^β, u^β) of (P^β) with $\beta \in I$. Moreover, if $u^0 = z$, then such a constant β_0 always exists and I can be chosen as any compact subinterval of $(0, \beta_0]$.

Using Lemmas 3.6 and 3.7 the proof of this proposition is quite analogous to that of Proposition 3.4. In the present case (K_1, K_2, K_3) is replaced by $(\sqrt{2}, 1/\pi, \sqrt{(1/\pi^2) + 1})$.

Special cases in which the coercivity estimate holds with $\beta = 0$ are quite similar to the multidimensional case and hence we will not explicitly state the analogue of Proposition 3.5 in dimension one.

4. Numerical results. In this section we briefly report on our practical experience with the augmented Lagrangian technique (2.8), (2.9) to determine q in (1.1) from data for u. We carried out extensive testing in dimensions one and two. These results will be presented in a forthcoming paper [8*] (see also [10], [11]) and we will therefore give only two typical examples.

Example 1. This is the problem of determining q in

$$-(qu_x)_x = f \quad \text{on } (0,1), \qquad u(0) = u(1) = 0 \tag{4.1}$$

where

$$f(x) = \begin{cases} (18x-6)\dfrac{3\pi}{2}\cos\dfrac{3\pi x}{2} + 18\left(\dfrac{3}{2} + \sin\dfrac{3\pi x}{2}\right) & \text{for } x \in [0, \tfrac{1}{3}), \\ 0 \quad \text{for } x \in (\tfrac{1}{3}, \tfrac{2}{3}), \\ (18x-12)\dfrac{3\pi}{2}\cos\dfrac{3\pi x}{2} + 18\left(\dfrac{3}{2} + \sin\dfrac{3\pi x}{2}\right) & \text{for } x \in [\tfrac{2}{3}, 1], \end{cases}$$

and the "true coefficient" q^* is given by

$$q^*(x) = \frac{3}{2} + \sin\frac{3\pi x}{2}.$$

The corresponding solution $u(q^*) = z$ of (4.1) is

$$u(q^*) = \begin{cases} -9x^2 + 6x & \text{for } x \in [0, \tfrac{1}{3}], \\ 1 \quad \text{for } x \in (\tfrac{1}{3}, \tfrac{2}{3}], \\ -9x^2 + 12x - 3 & \text{for } x \in (\tfrac{2}{3}, 1]. \end{cases}$$

With f and z specified, it is immediately clear that q is not unique within the class of positive H^1 functions that satisfy $u(q) = z$, since its value over the interval $S = (1/3, 2/3)$ does not effect the solution u there. On the other hand, with the techniques of [12] it can easily be argued that $u(q) = u(q^*)$, $q \in H^1$, $q^* \in H^1$ implies $q = q^*$ on $[0,1]\backslash S$. Thus we expect a different behavior of the algorithm over S than over the complement of S.

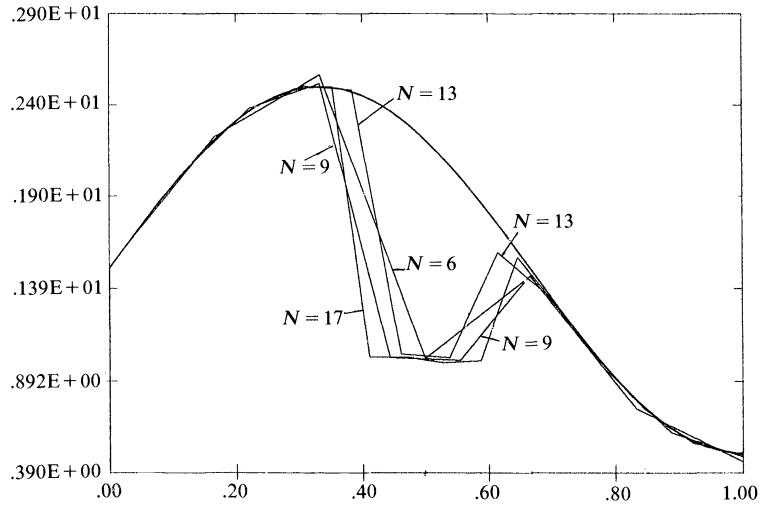
In Fig. 1 we give the numerical results for various values of N. Here N represents the index of discretization of the infinite-dimensional problem (2.8) by finite-dimensional ones involving linear spline subspaces H^N for the statespace H_0^1 and V^N for the parameter space H^1. More precisely we take

$$H^N = \operatorname{span}\{B_i^{2N}\}_{i=1}^{2N-1} \text{ and } V^N = \operatorname{span}\{B_j^N\}_{j=0}^N,$$

where B_i^N is the usual first-order B-spline on the interval $[0,1]$ corresponding to the mesh $\{x_k^N = k/N\}$, $k = 0, \cdots, N$:

$$B_k(x) = \begin{cases} N(x - x_{k-1}^N) & \text{for } x_{k-1}^N \leqq x \leqq x_k^N, \\ N(x_{k+1}^N - x) & \text{for } x_k^N \leqq x \leqq x_{k+1}^N, \\ 0 \quad \text{elsewhere} \end{cases}$$

where $x_{-1}^N = 0$, $x_{N+1}^N = 1$. Figure 1 gives the results after one iteration of the augmented Langrangian algorithm where $\lambda^1 = \mu^1 = 0$, $\gamma^2 = 100{,}000$ and $\beta = 0$. The start-up value for the minimization routine to solve (2.8) was taken as $(q^0, u^0) = (1, 0)$. The corresponding value for u^{17} is indistinguishable from $u(q^*)(= z)$ on all of $(0,1)$. For this example the use of the regularization term did not change the results significantly. In other examples with the same value for z, but with different values for q^* (and thus of f) the use of the regularization term decreased the L^2-error for $q^N - q^*$. The penalty



FUNCTION q METHOD ALG1
$n = 1$, $\gamma^2 = 100{,}000$, $N = 6, 9, 13, 17$, $\lambda_0 = 0$,
$\mu_0 = 0$, L^2-error = 4.134E−1, 5.164E−1, 4.286E−1, 5.514E−1

FIG. 1

parameters c_k were taken to be 1 for all k. The augmented Lagrangian algorithm for this example (as well as for the other examples that we tried) was quite insensitive to the choice of u^0 and q^0 as well as the choice of λ^1 and μ^1 and c_k. However, $\lambda^1 \neq 0$, $\mu^1 \neq 0$ increased the number of iterations that were required before convergence was obtained. We compared the augmented Lagrangian method with the output least squares approach and found that it is less sensitive with respect to the choice of the regularization parameter [10].

Example 2. Here we estimate q in

$$-(qu_x)_x - (qu_y)_y = f \quad \text{on } \Omega \tag{4.2}$$
$$u = 0 \quad \text{on } \Gamma$$

where $\Omega = [0, 1] \times [0, 1]$ and

$$f = 8\pi^2 \sin 2\pi x \sin 2\pi y(1 + 6x^2 y(1-y)) - 24\pi xy(1-y) \cos 2\pi x \sin 2\pi y$$
$$- 12\pi x^2(1-2y) \sin 2\pi x \cos 2\pi y.$$

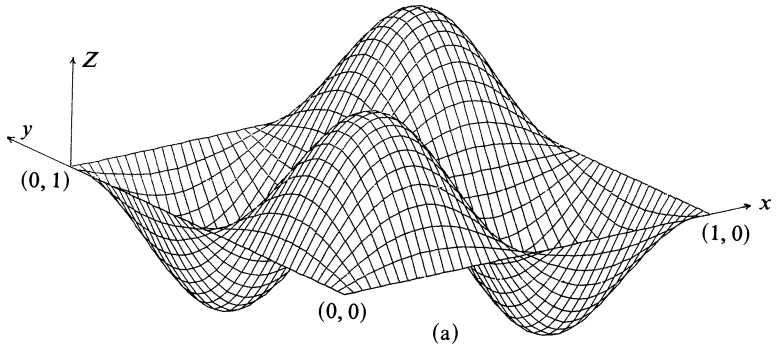
The true solution q^* is

$$q^* = 1 + 6x^2 y(1-y)$$

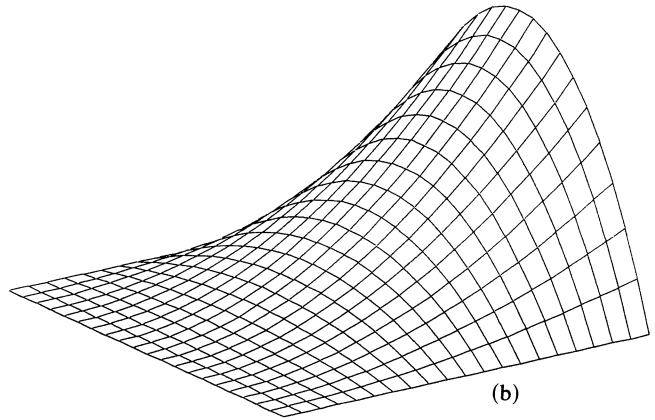
and the corresponding solution $u(q^*) = z$ for (4.2) is given by

$$u(q^*) = \sin 2\pi x \sin 2\pi y.$$

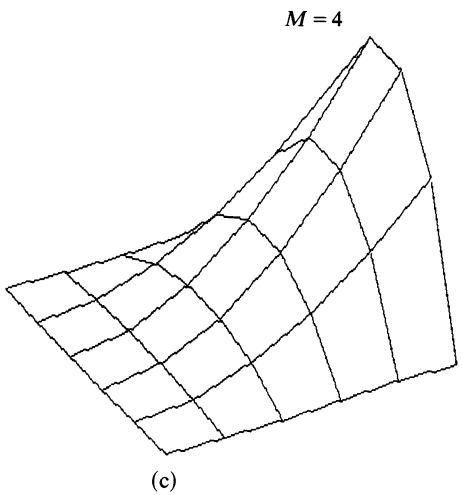
The discretization (4.2) is carried out by taking tensor linear spline subspaces $H^N \otimes H^N$ for the statespace H_0^1 [18], and tensor linear spline subspaces $V^N \otimes V^N$ for the coefficient space. Figures 2(a) and 2(b) give the graphs for z and q^*, respectively. The results after eight iterations of the augmented Lagrangian algorithm with $N = 5$ and $N = 9$ are given in Figs. 2(c) and 2(d). The results after just one iteration are essentially identical. These results are obtained with $\beta = \lambda^1 = \mu^1 = 0$ and $q_0 = 1$. We also carried out calculations where we assumed that only partial observations are available.

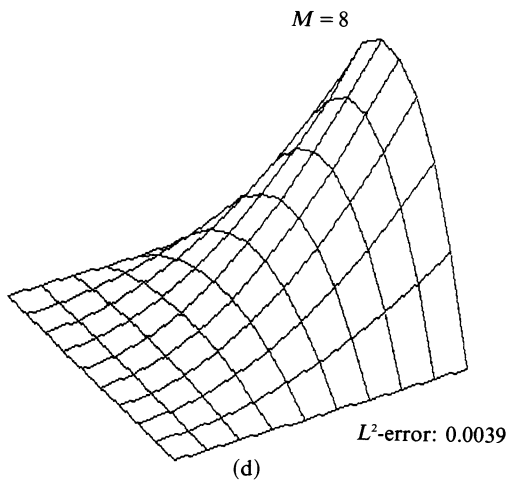


$Z = \sin 2\pi x \cdot \sin 2\pi y$
$q = 1 + 6x^2y(1-y)$



EXAMPLE 4. $q1$, $M = 4$, interpolated data at $(1/3{*}i, 1/3{*}j)$, $i, j = 0, 3$





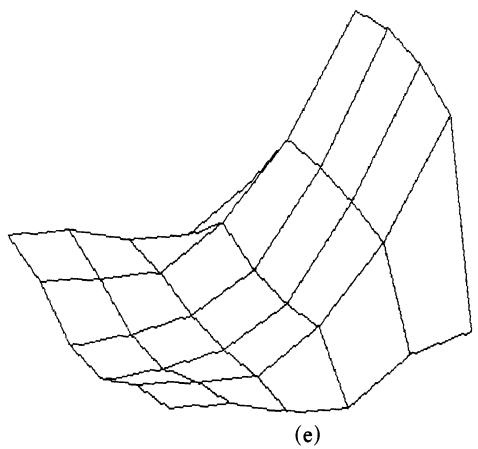


FIG. 2

Specifically, we took the values of $u(q^*)$ at the grid $[(.2i, .2j)\colon i, j = 0, \cdots, 5\}$ and calculated a bicubic interpolation z. Using this z in the augmented Lagrangian algorithm (2.8), (2.9) the resulting plot for q^5 is almost indistinguishable from Fig. 2(c). Then we tried the same procedure with data at $\{(i/3, j/3)\colon i, j = 0, \cdots, 3\}$ and the result for q^5 from these interpolated data is shown in Fig. 2(e).

Overall the augmented Lagrangian approach to estimate q in (1.1) proved to be very effective. This is especially true for the two-dimensional problem, where earlier experiments with the output least squares technique were not very encouraging numerically. Clearly, there is a wide variety of choices for implementing (2.8), (2.9). One variant of (2.8), (2.9) that proved to be effective numerically is the following (we specify it for $n = 2$ or 3).

Step 1. Choose $\lambda^1 = \mu^1 = 0$, $\{c_k\}_{k=1}^{\infty}$ monotonically increasing $c_k > c_0$.

Step 2. Put $k = 1$, $u_0 = z$.

Step 3. Determine q_k from

$$(P_{\text{equ}}) \text{ minimize } \frac{\beta}{2} N(q) + \langle \lambda^k, e(q, u_{k-1}) \rangle_{H_0^1} + \frac{c_k}{2} |e(q, u_{k-1})|_{H_0^1}^2$$

$$+ \mu^k \hat{g}(q, \mu^k, c_k) + \frac{c_k}{2} \hat{g}(q, \mu^k, c_k)^2$$

over $q \in H^2$ subject to $q \geqq \alpha$.

Step 4. Determine u_k from

$$(P_{\text{out}}) \underset{u \in H_0^1}{\text{minimize}} \frac{1}{2} |u - z|_{H_0^1}^2 + \langle \lambda^k, e(q_k, u) \rangle_{H_0^1} + \frac{c_k}{2} |e(q_k, u)|_{H_0^1}^2.$$

Step 5. $\lambda^{k+1} = \lambda^k + c_k e(q_k, u_k)$ and $\mu^{k+1} = \mu^k + c_k \hat{g}(q_k, \mu^k, c_k)$.

Step 6. If convergence is achieved, stop; otherwise put $k = k + 1$ and go to Step 3.

In our implementation for the two-dimensional problem we dropped the second-order derivative terms in the regularization functional of Step 3. This is partially because we used piecewise linear and piecewise constant functions to approximate q and hence second derivatives could only be taken approximately. Moreover, we expect that the second-order derivatives are only required analytically since we choose our coefficients q as H^2 functions, but we expect that this is not essential numerically.

Acknowledgment. We are grateful to Dr. M. Kroller who carried out the numerical calculations that we reported in this section.

REFERENCES

[1] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[2] G. CHAVENT, *Identification of distributed parameter systems: about the output least square method, its implementation and identifiability*, in Proc. 5th IFAC Symposium on Identification and System Parameter Estimation, Pergamon Press, New York, 1979, pp. 85-97.

[3] F. COLONIUS AND K. KUNISCH, *Stability for parameter estimation in two point boundary value problems*, J. Reine Angew. Math., 370 (1986), pp. 1-29.

[4] ———, *Output least squares stability in elliptic systems*, Appl. Math. Optim., 19 (1989), pp. 33-63.

[5] M. FORTIN AND R. GLOWINSKI, *Augmented Lagrangian Methods: Applications to Numerical Solutions of Boundary Value Problems*, North-Holland, Amsterdam, 1983.

[6] C. W. GROETSCH, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Research Notes in Mathematics 105, Pitman, Boston, 1984.
[7] M. R. Hestenes, *Optimization Theory, The Finite Dimensional Case*, John Wiley, New York, 1975.
[8] K. ITO AND K. KUNISCH, *The augmented Lagrangian method for equality and inequality constraints in Hilbert spaces*, Math. Programming, to appear.
[8*] K. ITO, M. KROLLER, AND K. KUNISCH, *A numerical study of the augmented Lagrangian method for the estimation of parameters in elliptic systems*, submitted.
[9] C. KRAVARIS AND J. H. SEINFELD, *Identification of parameters in distributed systems by regularization*, SIAM J. Control Optim., 23 (1985), pp. 217–241.
[10] M. KROLLER AND K. KUNISCH, *A numerical study of augmented Lagrangian method for the estimation of parameters in a two point boundary value problem*, Technical Report No. 87, Institute for Mathematics, Technical University of Graz, Graz, Austria, March 1987.
[11] ———, *A numerical study of an augmented Lagrangian method for the estimation of parameters in elliptic systems*, Technical Report, No. 101, Institute for Mathematics, Technical University of Graz, Graz, Austria, September 1987.
[12] K. KUNISCH, *Inherent identifiability of parameters in elliptic equations*, J. Math. Anal. Appl., 132 (1988), pp. 453–472.
[13] K. KUNISCH AND L. W. WHITE, *Regularity properties in parameter estimation of diffusion coefficients in elliptic boundary value problems*, Appl. Anal., 21 (1986), pp. 71–88.
[14] H. MAURER AND J. ZOWE, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.
[15] J. NEČAS, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson, Paris, 1967.
[16] V. T. POLYAK AND N. Y. TRET'YAKOV, *The method of penalty estimates for conditional extremum problems*, Z. Vychisl. Mat. I Mat. Fiz., 13 (1973), pp. 34–36.
[17] D. L. RUSSELL, *Some remarks on numerical aspects of coefficient identification in elliptic systems*, in Optimal Control of Partial Differential Equations, K. H. Hoffmann and W. Krabs, eds., Birkhäuser, Boston, 1984, pp. 210–228.
[18] M. H. SCHULZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
[19] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, New York, 1978.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 137-157, January 1990

007

OPTIMIZATION WITH AN AUXILIARY CONSTRAINT AND DECOMPOSITION*

GUY COHEN† AND BERNADETTE MIARA‡

Abstract. In the context of decomposition/coordination of a linear quadratic optimal control problem, Takahara's algorithm was an earlier version of the so-called Interaction Prediction Principle that can be examined in the more general framework of infinite-dimensional constrained optimization problems. This principle is both a *decomposition* principle and a *coordination* strategy based on a fixed point scheme. It has been later revisited in the general theory of the Auxiliary Problem Principle and the convergence of corresponding iterative algorithms has been analyzed. In this paper, we keep the same decomposition principle but we propose an alternative coordination strategy. The improvement brought by this new strategy is proved theoretically and illustrated by a numerical example. All of this is based on some manipulation of constrained optimization problems that we call the Auxiliary Constraint Principle.

Key words. optimization algorithms, decomposition, coordination, interaction prediction principle, convergence of algorithms

AMS(MOS) subject classifications. 49D27, 65K10, 93A15

1. Introduction.

The *Interaction Prediction Principle* (IPP) was first introduced by Takahara [1] (see also Mesarovic et al. [2]) in the context of decomposition/coordination of the classical linear quadratic (LQ) optimal control problem. Later on, Cohen [3], [4] gave a unified theory of decomposition/coordination algorithms in the framework of differentiable[1] mathematical programming in infinite-dimensional Hilbert spaces (in which of course deterministic optimal control problems can be casted). The IPP was shown to fall into the class of so-called *one-level algorithms* [3] for which conditions of convergence were given in [4]. The terminology "*one-level*" was chosen to outline the fact that this class of algorithms is based on a *fixed point* principle, that is, iterated values of primal and dual variables are directly exchanged between subproblems without the intervention of a coordination level.[2] This is in contrast with, e.g., *price coordination* [2], [8] where a coordinator iteratively updates the prices, generally by the Uzawa algorithm, which is indeed a *gradient* algorithm to maximize the dual (coordination objective) function.

At this point, we want to emphasize the difference between *gradient-like* or *variational* algorithms, the purpose of which is to minimize, maximize, or find the saddle-point of some (coordination) objective function, and *fixed point* strategies, the idea of which is to solve first-order optimality conditions. To make this point clearer, let us consider the simple problem of minimizing some real-valued differentiable convex functional f of two variables x and y belonging to the same Hilbert space or to spaces in duality ($\langle \cdot, \cdot \rangle$ denotes either the inner or the duality product). With these assump-

* Received by the editors February 3, 1988; accepted for publication (in revised form) March 1, 1989.
† Section Automatique, Ecole des Mines de Paris, 35 Rue Saint-Honoré, 77305 Fontainebleau Cedex, France. The author is also affiliated with the Institut National de Recherche en Informatique et Automatique Domaine de Voluceau, Rocquencourt, B.P. 105, 78150 Le Chesnay, France.

‡ Ecole Supérieure d'Ingénieurs en Electrotechnique et Electronique, 2 Boulevard Blaise Pascal, B.P. 99, 93162, Noisy-Le-Grand Cedex, France.

[1] This was later extended to nondifferentiable [5] and stochastic [6] optimization and to variational inequality [7] problems.

[2] Indeed, some coordination task can still be introduced through *under-* or *overrelaxation* strategies which may sometimes improve convergence.

tions, it is equivalent to say that we wish to solve the following system of equations:

$$\frac{\partial f(x, y)}{\partial x} = 0 \tag{1}$$

$$\frac{\partial f(x, y)}{\partial y} = 0. \tag{2}$$

A gradient algorithm uses the left-hand side current value of (1) (respectively, (2)) to update x (respectively, y). Suppose now that $f(x, y) = g(x, y) - \langle x, y \rangle$ so that (1), (2) are, respectively, equivalent to (3), (4) below

$$y = \frac{\partial g(x, y)}{\partial x} \tag{3}$$

$$x = \frac{\partial g(x, y)}{\partial y}, \tag{4}$$

which suggests the following fixed point (parallel) iterations:

$$y^{k+1} = \frac{\partial g(x^k, y^k)}{\partial x} \tag{5}$$

$$x^{k+1} = \frac{\partial g(x^k, y^k)}{\partial y}. \tag{6}$$

It should be noted that now (3) (that is equivalently (1)) is used to update y (and mutatis mutandis for x). Moreover, whereas with the gradient algorithm $f(x^k, y^k)$ should decrease, there is no reason why this should be the case with the fixed point algorithm. A convergence proof for the latter would rely upon an ad hoc Lyapounov function (or upon a contraction argument).

These considerations apply to the class of one-level algorithms as will be shown later on with more details. They largely explain the discrepancy that shows up in [4] between convergence conditions obtained for this class and those for all other classes of algorithms described there, which were of a variational nature. More specifically, there is no equivalent for these latter classes of the *prerequisite* conditions [4, (26-1) or (26-2)] imposed to prove convergence of one-level algorithms (these conditions will be recalled in § 2.4). In all other cases, conditions are only imposed to the step lengths used in these gradient-like algorithms under classical convexity and other technical conditions. Note also that Cohen was able to give a convergence proof for one-level algorithms in the context of linear equality constraints and quadratic cost functions only, whereas, for all the other algorithms, he was able to deal with general convex problems including inequality constraints as well.

The above remarks led us to reconsider the coordination strategy traditionally used in conjunction with the IPP. It is important to distinguish this *decomposition* principle, that is, a way to formulate independent subproblems using some coordination instruments, from the *coordination* strategy itself which is the way of updating these parameters from one iteration to the next. What we are going to do is modify the latter without altering the former. The new coordination strategy will be of a variational rather than of a fixed point nature. Consequently, we shall be able to avoid the prerequisite convergence condition mentioned above, which means that this new approach is more often applicable than the earlier one. Moreover, we shall be able to consider general convex cost functions, but we shall still be limited to affine equality constraints only for technical reasons that will be discussed later.

We first introduce a new way of manipulating a given constrained optimization problem, that is, a way of replacing it by an equivalent one which in this case will have more constraints and more variables. This may seem a rather odd idea, but the interest of this manipulation, which we call the *Auxiliary Constraint Principle* (ACP), will be more apparent later on. We even believe that this interest is broader than the application done here but we shall not elaborate on this. For the time being, let us say that when coupling between potential subproblems arises from constraints, either these constraints are relaxed by appealing to duality (this is what is done in the *Interaction Balance Principle* (IBP) [2] also known as *price decomposition* [8]) or, in one way or another, these coupling constraints must be replaced by uncoupled (auxiliary) ones. In § 2.3, we shall recall how this is achieved in the context of the *Auxiliary Problem Principle* (APP) by a proper choice of the auxiliary function when deriving one-level algorithms. With the ACP, the idea is different in that the auxiliary constraint is introduced by manipulation of the original problem and *before* one appeals to the APP in order to obtain iterative algorithms.

The rest of the paper is organized as follows. The next section is devoted to a summary of the situation starting with Takahara's algorithm later formalized by the IPP, followed by a brief description of the APP framework and the convergence conditions obtained for the so-called one-level algorithm which is directly connected to those earlier algorithms. The ACP is introduced in § 3 and a new coordination strategy is derived from it. Section 4 mixes the ACP and the APP to propose an alternative algorithm to one-level algorithms. A convergence theorem is stated but its proof is given in an appendix. A section is devoted to showing how Takahara's algorithm is modified using the new coordination strategy, and to giving an account of some simple numerical experiments, including a comparison of the traditional fixed point strategy with the new proposed strategy. In these experiments, we study the effect of increasing the interaction magnitude between subproblems. In the conclusion, we discuss open problems and topics of future research.

2. The IPP and its connection with the APP. In order for this paper to be reasonably self-contained, we first recall Takahara's algorithm in its original context of LQ optimal control problems [1]. This was the first appearance of the IPP later developed by Mesarovic and his coauthors [2]. We then abstract this algorithm in a more general context of constrained optimization. On the other hand, we recall the APP framework introduced by Cohen [3], [4], and we show how the IPP can be embedded in this framework (and somewhat extended). Finally, we recall the convergence conditions which result from these considerations.

2.1. Takahara's algorithm. Consider the classical LQ optimal control problem

$$\min \frac{1}{2}\int_0^T x^*Qx + u^*Ru\,dt \tag{7}$$

$$\dot{x} = Fx + Gu, \quad 0 \leqq t \leqq T, \quad x(0) = x_0 \text{ given}, \tag{8}$$

where the star denotes transposition. Moreover, assume that system (8) is made of N interconnected subsystems. With obvious notations, (8) can be rewritten for decomposition purposes as follows

$$i = 1, \cdots, N \begin{cases} F_{ii}x_i + G_{ii}u_i - v_i - \dot{x}_i = 0 \\ \sum_{j \neq i} (F_{ij}x_j + G_{ij}u_j) + v_i = 0 \end{cases} \tag{9}$$

where v_i is the ith interaction variable. In the same way, the necessary Pontryagin optimality conditions (which are also sufficient if Q—respectively R—is nonnegative

definite—respectively, definite positive—which we assume) can be globally written as

$$\begin{aligned} \dot{x} &= Fx + Gu, & x(0) &= x_0 \\ \dot{p} &= -F^*p - Qx, & p(T) &= 0 \\ 0 &= Ru + G^*p \end{aligned}$$

where p is the costate vector. The above conditions can be rewritten as follows

$$(10) \qquad i = 1, \cdots, N \begin{cases} \dot{x}_i = F_{ii}x_i + G_{ii}u_i - v_i \\ \dot{p}_i = -F_{ii}^*p_i - Q_{ii}x_i - \mu_i \\ 0 = R_{ii}u_i + G_{ii}^*p_i + \nu_i \\ v_i = -\sum_{j \neq i} (F_{ij}x_j + G_{ij}u_j) \\ \mu_i = \sum_{j \neq i} (F_{ij}^*p_j + Q_{ij}x_j) \\ \nu_i = \sum_{j \neq i} (G_{ij}^*p_j + R_{ij}u_j) \end{cases}$$

which can be reinterpreted as the optimality conditions of the following LQ problems:

$$(11) \qquad i = 1, \cdots, N \begin{cases} \min \int_0^T \frac{1}{2}(x_i^*Q_{ii}x_i + u_i^*R_{ii}u_i) + \mu_i^*x_i + \nu_i^*u_i \, dt \\ \dot{x}_i = F_{ii}x_i + G_{ii}u_i - v_i. \end{cases}$$

Takahara's algorithm consists of the following fixed point algorithm.

ALGORITHM 1 (Takahara). *At iteration k, knowing x^k, u^k, and p^k, compute the coordination parameters v_i^k, μ_i^k, and ν_i^k using the last three formulas* (10) *and produce the updated values x^{k+1}, u^{k+1}, and p^{k+1} by solving the subproblems* (11).

The fact that the interaction input v_i is "predicted" at some value to decouple the subsystems gave the name IPP to the decomposition method. However, there are also dual variables[3] μ_i and ν_i that must also be "predicted."

2.2. The IPP in the framework of constrained optimization. We now give a simpler view of the IPP by considering optimal control problems as particular instances of constrained optimization problems (in infinite-dimensional spaces). The decision variables are the pair of vector trajectories $\{x(t), u(t)\}_{0 \leq t \leq T}$ and the constraint is the dynamic equation (8) (whose multiplier is p). We shall not discuss the issue of proper topologies in which optimal control problems must be set and we refer the interested reader to [3], [9] for details. Hereafter, constrained optimization problems will be considered in Hilbert spaces only. From now on, let $\mathcal{U}$ denote the Hilbert space of decision variables and $\mathcal{C}$ denote that of constraints, let $J : \mathcal{U} \mapsto \mathbb{R}$ denote the cost function and $\Theta : \mathcal{U} \to \mathcal{C}$ denote the constraint function. We consider the constrained optimization problem

$$(12) \qquad \min J(u) \quad \text{subject to} \quad \Theta(u) = 0.$$

Introducing the Lagrange multiplier p which lies in the dual space $\mathcal{C}^*$ and the Lagrangian $L(u, p) = J(u) + \langle p, \Theta(u) \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the duality product between

[3] These variables should indeed be considered as dual despite the presence of Qx or Ru in their expressions because these latter expressions should be viewed as partial derivatives of the cost function with respect to x and u, respectively.

$\mathscr{C}$ and $\mathscr{C}^*$, under differentiability assumptions and a constraint qualification condition, a necessary condition for some u^* to be a solution of (12) is that there exists some p^* such that $L'_u(u^*, p^*) = 0$ and $L'_p(u^*, p^*) = 0$, where, e.g., L'_u denotes the partial derivative with respect to u. This yields

$$\begin{cases} J'_u(u^*) + [\Theta'_u(u^*)]^* p^* = 0 \\ \Theta(u^*) = 0 \end{cases} \tag{13}$$

where the star over an operator denotes its adjoint operator.

Suppose now that some decompositions of $\mathscr{U}$ and $\mathscr{C}$ into the product of N subspaces are given. Each component $\mathscr{C}_i$ is associated with a corresponding $\mathscr{U}_i$.[4] By J'_i we denote the partial derivative of J with respect to u_i, by Θ_i the mapping from $\mathscr{U}$ to $\mathscr{C}_i$ (that is, the composition of Θ and of the projection to $\mathscr{C}_i$), and finally by Θ'_{ij} the partial derivative of Θ_i with respect to u_j. Equations (13) can be rewritten by blocks, each block corresponding to a problem over the pair of spaces $(\mathscr{U}_i, \mathscr{C}_i^*)$. It is then easy to see that, in this new framework, Takahara's algorithm corresponds to a *relaxation* scheme over this set of equations. In what follows, (u_i^{k+1}, u_{-i}^k) is a shorter notation for $(u_1^k, \cdots, u_{i-1}^k, u_i^{k+1}, u_{i+1}^k, \cdots, u_N^k)$.

ALGORITHM 2 (IPP). *Knowing* (u^k, p^k), *compute* (u^{k+1}, p^{k+1}) *by solving*

$$i = 1, \cdots, N \begin{cases} J'_i(u_i^{k+1}, u_{-i}^k) + [\Theta'_{ii}(u_i^{k+1}, u_{-i}^k)]^* p_i^{k+1} + \sum_{j \neq i} [\Theta'_{ji}(u^k)]^* p_j^k = 0 \\ \Theta_i(u_i^{k+1}, u_{-i}^k) = 0. \end{cases} \tag{14}$$

These equations can again be interpreted as the necessary optimality conditions of the following subproblems

$$i = 1, \cdots, N \begin{cases} \min_{u_i} J_i(u_i, u_{-i}^k) + \sum_{j \neq i} \langle p_j^k, \Theta'_{ji}(u^k) \cdot u_i \rangle \\ \text{subject to } \Theta_i(u_i, u_{-i}^k) = 0. \end{cases} \tag{15}$$

Another version which differs from the above one—only if Θ is nonlinear—would amount to replacing the argument u^k by (u_i^{k+1}, u_{-i}^k) in the terms Θ'_{ji} of (14) and, correspondingly, replacing the last terms in the cost function of (15) by $\langle p_j^k, \Theta_j(u_i, u_{-i}^k) \rangle$.

2.3. The APP and one-level algorithms. Consider first a problem of the form

$$\min_{u \in U^f} J(u) + G(u) \tag{16}$$

where U^f, the feasible set, is a closed convex subset of $\mathscr{U}$, J and G are convex lower semicontinuous mappings from $\mathscr{U}$ to $\mathbb{R}$, and, moreover, J is differentiable. Let K be an auxiliary function of the same type as J. The APP leads to the following basic fixed point algorithm.

ALGORITHM 3 (APP). *Knowing* u^k, *compute* u^{k+1} *as the solution of*

$$\min_{u \in U^f} K(u) + \langle \varepsilon J'(u^k) - K'(u^k), u \rangle + \varepsilon G(u). \tag{17}$$

A justification of this fixed point strategy comes from the fact that if $u^k = u^{k+1}$ (i.e., u^k is a solution of (17)), then u^k is also a solution of (16) (this is proved by

[4] Indeed there is no necessity that the number of component subspaces $\mathscr{C}_i$ be exactly N—it may be smaller without trouble—but it will make our explanations simpler to assume that the two decompositions have the same cardinality.

writing the variational inequality associated to (17)). We refer the reader to [3], [4] for details and also for assumptions on K and conditions on the positive constant ε that guarantee convergence of the above algorithm. As far as decomposition is concerned, if some decomposition of $\mathcal{U}$ into $\mathcal{U}_1 \times \cdots \times \mathcal{U}_N$ is given, if $U^f = U_1^f \times \cdots \times U_N^f$ where $U_i^f \subset \mathcal{U}_i$ for $i = 1, \cdots, N$ (decoupled constraints), if G is additive (that is, $G(u) = \sum_i G_i(u_i)$), and finally if K is also chosen additive, then (17) splits into N independent subproblems.

This APP is generalized to saddle point problems to cope with coupling constraints. The idea is to formulate coupling constraints explicitly as equalities or inequalities and then to appeal to duality in order to relax these constraints. In this paper, we limit ourselves to equality constraints only and we consider (12) again.[5] To solve the saddle point problem of the Lagrangian L with respect to (u, p) in a decomposed way, we introduce an auxiliary saddle function $\Psi(u, p)$. However, the question arises of knowing whether we want to get a decomposition with respect to the variable u only (this is what occurs in price decomposition) or a decomposition with respect to both variables u and p. In the latter case, we have to speak of a decomposition of the space in which p lies, namely here $\mathcal{C}^*$. Hence we assume that a decomposition of $\mathcal{C}$ into N components is given, as we did in the previous section, and we adopt the same notations as previously regarding the components of Θ and Θ'.

Our final purpose is to recover algorithms that look like (15) above, that is, to get subproblems which are initially saddle point problems but which can eventually be interpreted as constrained minimization problems. This is obtained by making a proper choice of the auxiliary function Ψ, namely by giving it the form of a Lagrangian function. Therefore we set

$$\Psi(u, p) = K(u) + \langle p, \Omega(u) \rangle \tag{18}$$

where K is again an auxiliary function of the same type as J and Ω is of the same type as Θ. At iteration k, the auxiliary problem then consists of solving a saddle point problem for the above auxiliary function to which we must add linear corrections analogous to those appearing in (17), namely

$$\Psi(u, p) + \langle \varepsilon L'_u(u^k, p^k) - \Psi'_u(u^k, p^k), u \rangle + \langle \rho L'_p(u^k, p^k) - \Psi'_p(u^k, p^k), p \rangle \tag{19}$$

where ρ is, as ε, a positive constant. Since (19) is affine in p, it looks like a Lagrangian, hence its saddle point may be reinterpreted as an auxiliary constrained optimization problem. We then get the so-called *one-level* algorithm.

ALGORITHM 4 (ONE-LEVEL). *Knowing* (u^k, p^k), *compute* (u^{k+1}, p^{k+1}) *as the primal-dual solution of*

$$\begin{cases} \min\limits_u K(u) + \langle \varepsilon J'(u^k) - K'(u^k), u \rangle + \langle p^k, [\varepsilon \Theta'(u^k) - \Omega'(u^k)] \cdot u \rangle \\ \\ \textit{subject to } \Omega(u) + \rho \Theta(u^k) - \Omega(u^k) = 0. \end{cases} \tag{20}$$

In order for this problem to split into N independent subproblems, one has to choose an *additive* K and a *block-diagonal* Ω. "Block-diagonal" of course means that Ω'_{ij} is identically null whenever $j \neq i$. Note that this condition is necessary and sufficient for the auxiliary function (18) to be additive with respect to the pair (u, p). Now, to precisely recover the subproblems (15), the auxiliary functions K and Ω must depend on the iteration index k, which does not conceptually make any problem, and the

[5] We could have kept uncoupled constraints $u \in U^f$ and an additive part G of the cost function as well.

components K_i^k and Θ_i^k (with obvious notations) must be chosen as follows whereas ε and ρ are set equal to one[6]

$$i = 1, \cdots, N \begin{cases} K_i^k(u_i) = J(u_i, u_{-i}^k) \\ \Omega_i^k(u_i) = \Theta_i(u_i, u_{-i}^k). \end{cases}$$

2.4. Convergence conditions of one-level algorithms. A convergence proof for Algorithm 4 was given only in the case of J and K quadratic and Θ and Ω affine [4]. We recall these convergence conditions.

THEOREM 1. *Let* $J(u) = \frac{1}{2}\langle u, Au\rangle + \langle b, u\rangle$, $K(u) = \frac{1}{2}\langle u, Bu\rangle$, $\Theta(u) = Du - d$, $\Omega(u) = Eu$, *where* A *and* B *are self-adjoint strongly monotone operators over* $\mathcal{U}$ *and* D *and* E *are linear onto operators from* $\mathcal{U}$ *to* $\mathcal{C}$. *Under the following condition (where "strongly monotone" is symbolically denoted by* >0*)*

$$DA^{-1}E^* + EA^{-1}D^* - DA^{-1}BA^{-1}D^* > 0 \text{ over } \mathcal{C}^* \tag{21}$$

it is possible to choose $\rho = \varepsilon$ *and the latter so that*

$$\begin{gathered} B - \varepsilon A/2 > 0 \quad \text{over } \mathcal{U} \\ DA^{-1}E^* + EA^{-1}D^* - DA^{-1}(B + \varepsilon A/2)A^{-1}D^* > 0 \quad \text{over } \mathcal{C}^* \end{gathered} \tag{22}$$

and then Algorithm 4 *converges to the unique solution* (u^*, p^*) *of* (12).

Once (21) is met, (22) can be satisfied for ε small enough. The prerequisite condition (21) can be further simplified into

$$DA^{-1}E^* + EA^{-1}D^* > 0 \quad \text{over } \mathcal{C}^* \tag{23}$$

for, if this latter condition is satisfied, the former can also be satisfied by changing E into αE with α large enough. Hence the problem is to find some auxiliary constraint operator E which is at the same time *block-diagonal* (for decomposition purposes) and which satisfies (23). Indeed, we do not know the answer to the question of existence of such an E in general, but it is clear that this depends heavily on the way the constraint space $\mathcal{C}$ is decomposed into component subspaces (that is, how constraints are grouped in blocks), and then on how these blocks are numbered (that is how subspaces $\mathcal{C}_i$ are paired with subspaces $\mathcal{U}_i$ or else how these blocks of constraints are allocated to subproblems in u_i). This issue is further discussed in [4, Remark 5.1]. However, these considerations will no longer be relevant when using the new coordination strategy introduced hereafter since we will be able to avoid condition (23) to prove convergence.

3. The ACP and a new coordination strategy. As we have just seen, coupling constraints must be handled by appealing to duality. But there are two ways of doing this. In price coordination[7] [8], these constraints no longer appear as constraints in the subproblems but they appear as additional terms in the subproblem objective functions. On the contrary, in the IPP approach, these constraints still appear at the subproblem level as constraints, but of course in a decoupled manner (see (15)), and the additional terms in the subproblem cost functions account for the coupling part

[6] Note that this common value of the convergence parameters ε and ρ may be outside the range of allowed values resulting from the convergence conditions recalled hereafter. But we never claimed that Algorithm 2 would always converge. As far as Takahara's algorithm is concerned, the proof given by the author in [1] was very involved and his convergence conditions were not very clear.

[7] This is also called IBP [2], which is embedded in the so-called family of *two-level* algorithms [3], [4].

only. Alternately in the APP/one-level approach (see (20)), constraints that appear in the subproblems are "auxiliary" constraints, and additional terms in the subproblem cost functions account for the discrepancy between original and auxiliary constraints. These auxiliary constraints, or more exactly the auxiliary operator Ω (or E), have been introduced through the choice of an auxiliary function (18) having a Lagrangian form. Now we are going to introduce this auxiliary operator Ω directly at the level of an equivalent formulation of the original problem (12), that is *prior* to any recourse to the APP, the aim of the latter being to define an iterative algorithm to compute a solution of (12). This way of formulating an equivalent problem to a given constrained optimization problem is referred to as the *Auxiliary Constraint Principle* (ACP).

3.1. The ACP.

THEOREM 2. *In addition to* (12) *and to its Lagrangian introduced earlier, consider the following problem*

$$\text{(24)} \qquad \begin{cases} \min\limits_{u,v} J(u) \\ \text{subject to } \Omega(u) - v = 0 \\ \qquad\qquad \Theta(u) + v - \Omega(u) = 0 \end{cases}$$

(*where v is a new decision variable in $\mathscr{C}$ and Ω maps $\mathscr{U}$ into $\mathscr{C}$*) *and the corresponding Lagrangian*

$$\text{(25)} \qquad \mathscr{L}(u, v; p, q) = J(u) + \langle p, \Omega(u) - v\rangle + \langle q, \Theta(u) + v - \Omega(u)\rangle.$$

Then problems (12) *and* (24) *are equivalent in the following sense*:

(i) *If $(u^*, v^*; p^* q^*)$ is a saddle point of $\mathscr{L}$ over $(\mathscr{U} \times \mathscr{C}) \times (\mathscr{C}^* \times \mathscr{C}^*)$, then $p^* = q^*$ (and of course $v^* = \Omega(u^*)$) and (u^*, p^*) is a saddle point of L.*

(ii) *Conversely, if (u^*, p^*) is a saddle point of L, then $(u^*, \Omega(u^*), p^*, p^*)$ is a saddle point of $\mathscr{L}$.*

Proof.

(i) From the right-hand inequality of the saddle point of $\mathscr{L}$, it is easy to conclude that necessarily $p^* = q^*$ and, of course, from the left-hand inequality that $v^* = \Omega(u^*)$, and then these inequalities reduce to those of the saddle point of L.

(ii) The proof of the converse statement is straightforward. □

Remark 1. This theorem appeals to the *global* theory of duality in that it speaks of saddle points of Lagrangian. Existence of a Lagrangian saddle point is a *sufficient* condition for existence of a solution to problems such as (12) or (24). There are also *necessary* first-order optimality conditions in the framework of a *local* theory of duality (involving Khun-Tucker multipliers in the differentiable case). Obviously, it is possible to state a result similar to Theorem 2 in this context.

3.2. A new coordination strategy for the IPP. Since we consider equality constraints only, and since we wish to remain in the context of convex programming, we deal only with affine constraints and we set $\Theta(u) = Du - d$ and $\Omega(u) = Eu$ as we did in Theorem 1.[8] The stationarity conditions for $\mathscr{L}$ can be expressed in the following way:

$$\text{(26)} \qquad \mathscr{L}'_u(u^*, v^*; p^*, q^*) = 0 \Leftrightarrow J'(u^*) + (D^* - E^*)q^* + E^* p^* = 0$$

$$\text{(27)} \qquad \mathscr{L}'_p(u^*, v^*; p^*, q^*) = 0 \Leftrightarrow Eu^* - v^* = 0$$

$$\text{(28)} \qquad \mathscr{L}'_v(u^*, v^*; p^*, q^*) = 0 \Leftrightarrow q^* - p^* = 0$$

[8] With Algorithm 4, it does not matter if we rather set $\Omega(u) = Eu - d$, as it does not matter if we add a linear function to K. With (24) or (30), that amounts to changing v into $v - d$.

$$\mathscr{L}'_q(u^*, v^*; p^*, q^*) = 0 \Leftrightarrow (D-E)u^* + v^* - d = 0. \tag{29}$$

As far as decomposition is concerned, we shall again assume that decompositions of $\mathscr{U}$ and $\mathscr{C}$ into subspaces are given and that E is block-diagonal with respect to these decompositions. To further simplify this introductory discussion, let us temporarily assume that J is additive. The decomposition idea behind the IPP can be viewed as that of dealing with u and p at the lower (subsystem) level and with v and q at the upper (coordination) level. Indeed, with our assumptions, and if v and q are fixed at some values, say v^k and q^k, the task of solving (26)-(27) splits into N independent tasks which can be interpreted as those of solving the following subproblems

$$i = 1, \cdots, N \begin{cases} \min\limits_{u_i} J_i(u_i) + \sum\limits_{j \neq i} \langle q_j^k, D_{ji}u_i \rangle + \langle q_i^k, (D_{ii} - E_i)u_i \rangle \\ \text{subject to } E_i u_i = v_i^k. \end{cases} \tag{30}$$

Let (u^{k+1}, p^{k+1}) denote an optimal solution. Then, the fixed point coordination strategy usually associated with the IPP uses (28) to update q, namely by setting $q^{k+1} = p^{k+1}$, and (29) to update v, that is, $v^{k+1} = (E-D)u^{k+1} + d$. This strategy may be improved by introducing under- or overrelaxation (according to whether the ρ_i's hereafter are smaller or larger than one)

$$\begin{aligned} v^{k+1} &= (1-\rho_1)v^k + \rho_1((E-D)u^{k+1} + d) \\ &= v^k - \rho_1(Du^{k+1} - d) \end{aligned} \tag{31}$$

$$\begin{aligned} q^{k+1} &= (1-\rho_2)q^k + \rho_2 p^{k+1} \\ &= q^k + \rho_2(p^{k+1} - q^k). \end{aligned} \tag{32}$$

At this point, the reader should remember our discussion of the introduction around (1)-(6). It results from that discussion that we can imagine the following alternative coordination strategy of a gradient or variational nature.

ALGORITHM 5 (ACP). *Knowing* (v^k, q^k), *compute* (u^{k+1}, p^{k+1}) *as a solution of* (30) *and update* (v, q) *by the gradient formulas*

$$v^{k+1} = v^k - \rho_1 \mathscr{L}'_v(u^{k+1}, v^k; p^{k+1}, q^k) = v^k - \rho_1(q^k - p^{k+1}) \tag{33}$$

$$q^{k+1} = q^k + \rho_2 \mathscr{L}'_q(u^{k+1}, v^k; p^{k+1}, q^k) = q^k + \rho_2(Du^{k+1} - d), \tag{34}$$

where ρ_1 *and* ρ_2 *are positive step lengths.*

The comparison of (31), (32) with (33), (34) reveals the important difference between these two strategies.

Remark 2. We have used the fact that $Eu^{k+1} = v^k$ (see (30)) in both (31) and (34). However, it is important to keep the term $-\langle q_i^k, E_i u_i \rangle$ in the cost function (30) even though the value of $E_i u_i$ is a priori known, for, otherwise, the value of p^{k+1} would be changed, and p^{k+1} is used in both (32) and (33).

3.3. The new strategy is a parallel Arrow–Hurwicz algorithm. A justification of the new coordination strategy comes from the study of the following functional

$$\Lambda(v, q) := \inf_u \sup_p \mathscr{L}(u, v; p, q). \tag{35}$$

Note that problem (24) is equivalent to finding $\inf_{u,v} \sup_{p,q} \mathscr{L}(u, v; p, q)$ (without any convexity assumption—this is a classical result in duality theory). Assuming that $\sup_q$ can be inverted with $\inf_u$, the problem reduces to finding $\inf_v \sup_q \Lambda(v, q)$ or the saddle

point of Λ if such a saddle point exists. Then Algorithm 5 is nothing but the Arrow-Hurwicz algorithm [10] in its parallel (rather than sequential) version in that v and q are updated at the same time instead of sequentially. All these facts are precisely stated in the following theorem.

THEOREM 3.

(i) *If there exists a saddle point* $(u^*, v^*; p^*, q^*)$ *of* $\mathscr{L}$, *then* (v^*, q^*) *is a saddle point of* Λ.

(ii) *If* J *is convex and if* Θ *and* Ω *are affine, then* Λ *is a convex-concave (or saddle) function.*

(iii) *If in the definition* (35) *of* Λ, *the* inf sup *is indeed a saddle point, and if the argument* $(\hat{u}(v, q), \hat{p}(v, q))$ *of this saddle point is unique, then* Λ *is differentiable and we have*

$$\begin{aligned}\Lambda'_v(v, q) &= \mathscr{L}'_v(\hat{u}(v, q), v; \hat{p}(v, q), q)\\ &= q - \hat{p}(v, q)\\ \Lambda'_q(v, q) &= \mathscr{L}'_q(\hat{u}(v, q), v; \hat{p}(v, q), q)\\ &= \Theta(\hat{u}(v, q)) + v - \Omega(\hat{u}(v, q)).\end{aligned}$$

Proof.

(i) We have the following general inequalities

$$\begin{aligned}\inf_{u,v}\sup_{p,q}\mathscr{L} &\geqq \inf_{v}\sup_{q}\inf_{u}\sup_{p}\mathscr{L}\\ &\geqq \sup_{q}\inf_{u,v}\sup_{p}\mathscr{L}\\ &\geqq \sup_{p,q}\inf_{u,v}\mathscr{L}.\end{aligned}$$

Since it is assumed that $\mathscr{L}$ does have a saddle point, the two extreme sides are equal. Then, the equality of the two inner sides means that Λ has a saddle point. The rest of the statement is easy to prove.

(ii) Let us introduce the intermediate functional

$$\lambda(u, v, q) := \sup_{p} \mathscr{L}(u, v; p, q).$$

Since $\mathscr{L}$ is *jointly* convex in (u, v), λ is also (jointly) convex in (u, v) as the upper hull of a family of convex functions. Since $\mathscr{L}$ is *jointly* concave in (p, q), λ is concave in q. This is a general result: if a function $f(x, y)$ is *jointly* concave (respectively, convex) in (x, y), the function $\varphi(y) := \sup_x f(x, y)$ (respectively, $\inf_x f(x, y)$) is concave (respectively, convex) in y. Also, for this latter reason, $\Lambda(v, q) = \inf_u \lambda(u, v, q)$ is convex in v. Finally, as the lower hull of a family of concave functions, it is concave in q.

(iii) This is a generalization of a result by Danskin [11] which states a similar result for either of the two functions $\varphi(y)$ introduced above. We skip the proof of this generalization. □

4. Decomposition/coordination algorithms derived from the APP and the ACP. In the same way as Algorithm 4 is an extension of Algorithm 2, we are going to propose an extension of Algorithm 5 above using the APP. We also propose several variants. We then study convergence issues.

4.1. General algorithms. The problem is to compute the saddle point of $\mathscr{L}$. Instead of (18), we choose the following auxiliary function[9]

$$\Phi(u, v; p, q) = K(u) + \frac{1}{2}\|v\|^2 - \frac{1}{2}\|q\|^2 \tag{36}$$

and we make the same kind of linear corrections as in (19). The way we put the superscripts k and $k+1$ hereafter translates the fact that we want to deal with (u, p) at the lower level and *then* with (v, q) at the upper level. Notice also that Φ does not directly depend on p. Actually, the term $\mathscr{N}(u, p) = \langle p, Eu\rangle$ of $\mathscr{L}$ should be considered as "additive" since E will be chosen block-diagonal and (u_i, p_i) will be dealt with simultaneously. Therefore we deal with $\mathscr{N}(u, p)$ as we did with the term $G(u)$ when obtaining (17). We set $\mathscr{M}(u, v; p, q) = \mathscr{L}(u, v; p, q) - \mathscr{N}(u, p)$. Finally, this leads us to calculate

$$\begin{cases} \Phi(u, v; p, q) + \varepsilon\mathscr{N}(u, p) \\ \qquad + \langle \varepsilon\mathscr{M}'_u(u^k, v^k; p^k, q^k) - \Phi'_u(u^k, v^k; p^k, q^k), u\rangle \\ \qquad + \langle \varepsilon\mathscr{M}'_p(u^k, v^k; p^k, q^k) - \Phi'_p(u^k, v^k; p^k, q^k), p\rangle \\ \qquad + \langle \rho_1\mathscr{M}'_v(u^{k+1}, v^k; p^{k+1}, q^k) - \Phi'_v(u^{k+1}, v^k; p^{k+1}, q^k), v\rangle \\ \qquad + \langle \rho_2\mathscr{M}'_q(u^{k+1}, v^k; p^{k+1}, q^k) - \Phi'_q(u^{k+1}, v^k; p^{k+1}, q^k), q\rangle. \end{cases} \tag{37}$$

The algorithm consists of computing the $\max_p \min_u$ of this expression, yielding (u^{k+1}, p^{k+1}), which can be interpreted as the solution of a constrained optimization problem, and then computing (v^{k+1}, q^{k+1}) by solving for the $\max_q \min_v$ which yields (33)-(34). We summarize this as follows.

ALGORITHM 6 (ACP+APP−PARALLEL). *Knowing* (u^k, v^k, p^k, q^k), *compute* (u^{k+1}, p^{k+1}) *as the solution of the following auxiliary problem*

$$\begin{cases} \min\limits_u K(u) + \langle \varepsilon J'(u^k) - K'(u^k), u\rangle + \varepsilon\langle q^k, (D-E)u\rangle \\ \text{subject to } \varepsilon(Eu - v^k) = 0 \end{cases} \tag{38}$$

and then update (v, q) *using formulas* (33)-(34).

Remark 3. It is important not to forget ε in factor of the constraint of (38) since otherwise p^{k+1} used in (33) would be differently scaled. Along the same line, we recall the second part of Remark 2 which still applies here. Finally, comparing (20) and (38), it is seen that Eu in (38) must not be identified with $\Omega(u)$ in (20) but rather with $\varepsilon\Omega(u)$.

As pointed out in Remark 2, we have used the fact that $Eu^{k+1} = v^k$ to simplify the update formula (34) of q.

From the decomposition point of view, whenever decompositions of $\mathscr{U}$ and $\mathscr{C}$ are given, if K—respectively, E—is chosen additive—respectively, block-diagonal—with respect to this (these) decomposition(s), then (38) splits into N independent subproblems.

We can imagine several other versions of the above algorithm. One corresponds to a *sequential* rather than a *parallel* algorithm in v and q, v being updated before q: this amounts to replacing v^k by v^{k+1} in the last term of (37).

[9] This is the usual way to get a gradient algorithm for v and q.

ALGORITHM 7 (ACP+APP−SEQUENTIAL−v BEFORE q). *In Algorithm* 6, *replace the update formula* (34) *for* q *by*

$$q^{k+1} = q^k + \rho_2((D-E)u^{k+1} + v^{k+1} - d). \tag{39}$$

Another sequential version is obtained by updating q before v: this amounts to replacing q^k by q^{k+1} in the next to last term of (37).

ALGORITHM 8 (ACP+APP−SEQUENTIAL−q BEFORE v). *In Algorithm* 6, *replace the update formula* (33) *for* v *by*

$$v^{k+1} = v^k + \rho_1(p^{k+1} - q^{k+1}). \tag{40}$$

See § 5.2 for a comparison of these two sequential versions.

Finally, we can imagine an *implicit* version which results from both modifications indicated above.

ALGORITHM 9 (ACP+APP−IMPLICIT). *In Algorithm* 6, *replace the update formulas* (33)-(34) *for* (v, q) *by the implicit formulas* (40) *and* (39) *which are equivalent to the following explicit form*

$$v^{k+1} = v^k + \frac{\rho_1}{1+\rho_1\rho_2}(p^{k+1} - q^k - \rho_2(Du^{k+1} - d)) \tag{41}$$

$$q^{k+1} = q^k + \frac{\rho_2}{1+\rho_1\rho_2}(Du^{k+1} - d + \rho_1(p^{k+1} - q^k)). \tag{42}$$

Indeed, there is another way to get this implicit algorithm directly. Before we explain this let us make the following remark.

Remark 4. It is equivalent to introduce ε, ρ_1 and ρ_2 as we did in (37), or to set $\varepsilon = \rho_1 = \rho_2 = 1$ in (37) but to redefine Φ (see (36)) as

$$\Phi(u, v; p, q) = \frac{1}{\varepsilon}K(u) + \frac{1}{2\rho_1}\|v\|^2 - \frac{1}{2\rho_2}\|q\|^2. \tag{43}$$

To get the parallel version, we defined $\mathcal{N}(u, p)$ as $\langle p, Eu\rangle$ on the basis of the fact that u and p were dealt with at the same level and thus a nonadditive term appearing in $\mathcal{L}$ could be preserved in the auxiliary problem. The same kind of reason can now be advocated for the pair (v, q) and the term $\langle q, v\rangle$ appearing in $\mathcal{L}$ (see (25)). Therefore, redefining $\mathcal{N}(u, v; p, q)$ as $\langle p, Eu\rangle + \langle q, v\rangle$ and again $\mathcal{M}$ as the complement of $\mathcal{N}$ to $\mathcal{L}$, we can use this observation in conjunction with Remark 4.1 above to get the implicit version directly from (37).

4.2. Convergence.

DEFINITION 1. We say that J is *strongly convex with constant* a if and only if $\exists a > 0$: $\forall u$, $\bar{u}$ and $\alpha \in [0, 1]$,

$$J(\alpha u + (1-\alpha)\bar{u}) \leqq \alpha J(u) + (1-\alpha)J(\bar{u}) - \frac{a}{2}\alpha(1-\alpha)\|u - \bar{u}\|^2. \tag{44}$$

If J is differentiable, it is equivalent to say that J' is strongly monotone with modulus a. Also, this property implies that

$$J(\bar{u}) \geqq J(u) + \langle J'(u), \bar{u} - u\rangle + \frac{a}{2}\|u - \bar{u}\|^2. \tag{45}$$

The Lipschitz property of J' with constant A, namely

$$\exists A > 0 : \|J'(\bar{u}) - J'(u)\| \leqq A\|\bar{u} - u\| \tag{46}$$

implies that

$$J(\bar{u}) \leqq J(u) + \langle J'(u), \bar{u} - u\rangle + \frac{A}{2}\|u - \bar{u}\|^2. \tag{47}$$

THEOREM 4. *We make the foliowing assumptions*:

- *J (respectively, K) is strongly convex with constant a (respectively, b) and differentiable with a Lipschitz derivative with constant A (respectively, B).*
- *D and E are linear continuous operators and E is onto.*

Then there exist a unique primal solution u^* to (12), *and a unique primal-dual solution (u^{k+1}, p^{k+1}) to* (38). *Moreover if ε is chosen in the open interval $(0, b/A)$, there exists an interval $(0, \bar{\rho}_1(\varepsilon))$ in which ρ_1 must be chosen, and then an interval $(0, \bar{\rho}_2(\varepsilon, \rho_1))$ in which ρ_2 must be chosen so that Algorithm 6 generates a sequence $\{u^k\}$ which converges to u^* (and consequently, $\{v^k\}$ converges to Eu^*). The sequences $\{p^k\}$ and $\{q^k\}$ are bounded and have the same cluster points (in the weak topology). Any such cluster point forms with u^* a saddle point of L. Finally, if D is also assumed to be onto, the dual solution p^* of* (12) *is unique and (p^k, q^k) (strongly) converges to (p^*, q^*).*

Comments. The exact expressions of the bounds $\bar{\rho}_1(\varepsilon)$ and $\bar{\rho}_2(\varepsilon, \rho_1)$ would be very involved as shown by the proof given in the appendix, but we can say that if ε tends to either end of its allowed interval, then $\bar{\rho}_1(\varepsilon)$ tends to 0. The same occurs to $\bar{\rho}_2(\varepsilon, \rho_1)$ if ε or ρ_1 approaches its bounds.

Concerning the sequential and the implicit versions (Algorithms 7, 8 and 9), essentially the same convergence theorems as the above can be stated and similarly proved.

5. Applications and numerical results. Before discussing the results of very simple numerical experiments, let us come back for a while to Takahara's algorithm and show how the new coordination strategy looks like in this particular instance.

5.1. Takahara's algorithm revisited. We are not going to make all the calculations, leaving them as an exercise to the reader, but we will only give the main indications. First of all, the variable u of the general theory must be identified with the pair (u, x) of § 2.1, $Du - d = 0$ is (8) (in the form $Fx + Gu - \dot{x} = 0$ with its given initial condition x_0), thus p is p, and of course $J(u)$ is (7). It is natural to choose $K(u)$ as the same integral cost but which retains only the block-diagonal part of Q and R. Finally, $Eu - v = 0$ and $Du - d + v - Eu = 0$ are the pair of equations (9) in that form.

Because of the connection of K with J, it is convenient to set $\varepsilon = 1$ (as we actually did in (30)). But it is not clear that the condition of Theorem 4, namely $\varepsilon < b/A$, does allow this value. Actually, it suffices that b be large enough so that $b > A$. This may be obtained by changing our previous choice of K, adding to it a term of the following form

$$\frac{\gamma}{2}\int_0^T \|u\|^2 + \|x\|^2\, dt \tag{48}$$

with γ large enough. With this at hand, subproblems (11) are unchanged,[10] except for the consequence of (48) which add terms $\gamma/2\int_0^T \|u - u^k\|^2 + \|x - x^k\|^2\, dt$ in the cost

[10] At stage k, set $v_i = v_i^k$ and likewise for μ_i and ν_i.

functions. Let $(u^{k+1}, x^{k+1}, p^{k+1})$ be the resulting solution. Then the new coordination strategy (Algorithm 6) updates μ_i^k and ν_i^k in the following way

$$i=1,\cdots,N\begin{cases} v_i^{k+1}=v_i^k+\rho_1(p_i^{k+1}-q_i^k) \\ q_i^{k+1}=q_i^k+\rho_2\left[\sum_{j\neq i}(F_{ij}x_j^{k+1}+G_{ij}u_j^{k+1})+v_i^k\right] \\ \mu_i^{k+1}=\sum_{j\neq i}(F_{ij}^*q_j^{k+1}+Q_{ij}x_j^{k+1}) \\ \nu_i^{k+1}=\sum_{j\neq i}(G_{ij}^*q_j^{k+1}+R_{ij}u_j^{k+1}). \end{cases}$$

With Algorithm 7, we must change v_i^k in the right-hand side of the second equation above into v_i^{k+1}. We let the reader imagine the other versions (Algorithms 8 and 9).

5.2. Numerical example. The following example is purposedly very simple so that it is possible to have a measure of performance which is independent of the initial conditions of iterations for the various algorithms we are going to compare. This is important since the dimensionality of the "state vector" of the iterations (the number of required initial conditions) is not the same for all of them. We consider quadratic cost functions and linear constraints so that the iterations can be put into the form of a linear system $z^{k+1}=Pz^k$ where z is the state vector, and P is a matrix depending on the convergence parameters ε, and (ρ_1, ρ_2) when appropriate. For all algorithms, we retained the best value of τ we have been able to obtain by the choice of convergence parameters as the measure of performance, where τ is the maximal modulus of eigenvalues of P.

We consider the following problem[11]

$$\min \frac{1}{2}\sum_{i=1}^{2}[\|x_i\|^2+\|u_i\|^2]$$

$$\text{subject to } u_i+\sigma x_j-x_i=0, \quad i,j=1,2, \quad j\neq i$$

where σ will serve us to make the interaction magnitude between the two subsystems more or less important. With respect to the general theory, $u=(u_1, u_2)$ must be identified with $((u_1, x_1), (u_2, x_2))$. We choose $K=J$ (as far as J is additive) and E as the block-diagonal part of D, that is

$$D=\begin{pmatrix}1 & -1 & 0 & \sigma\\ 0 & \sigma & 1 & -1\end{pmatrix} \quad \text{and} \quad E=\begin{pmatrix}1 & -1 & 0 & 0\\ 0 & 0 & 1 & -1\end{pmatrix}.$$

Notice that

$$DA^{-1}E^*+EA^{-1}D^*=\begin{pmatrix}4 & -2\sigma\\ -2\sigma & 4\end{pmatrix}$$

so that the condition (23) is not satisfied for $|\sigma|\geqq 2$.

For this example, we considered Algorithm 4 (fixed point coordination, without under- or overrelaxation scheme since this did not seem to improve things in this instance), Algorithm 6 (parallel Arrow-Hurwicz coordination strategy) and its sequential and implicit versions (Algorithms 7, 8, and 9). To save space, we do not give the expressions of matrix P for all these algorithms; they result from a straightforward application of the general theory. Let us just notice that the state vector z is 8-

[11] The solution is zero for all primal and dual variables but of course this does not matter as far as convergence is concerned.

dimensional with Algorithms 6, 7, 8, and 9 (z is the column vector whose entries are $(u_i, x_i, v_i, q_i)_{i=1,2}$)[12] whereas it is only 6-dimensional with Algorithm 4 (z is $(u_i, x_i, p_i)_{i=1,2}$). The numerical results summarized in Table 1 have been obtained using MATLAB on a Macintosh.

Several comments are in order. First, as predicted by theory, Algorithms 6–9 always converge, even for large interactions, although more and more slowly as interaction magnitude increases, whereas Algorithm 4 fails to converge for $\sigma \geqq 2$. Second, as it might also be expected, the sequential versions appear to be faster than the parallel one and the implicit version is the best of all.[13] Third, the same performance is obtained with both sequential versions whose results have been displayed in the same column of the table. We shall come back to this point later on. Finally, the "one-level" algorithm appears to be better than the other algorithms for small or moderate interaction magnitude, except for the implicit version, at least for the interaction magnitudes investigated above. This confirms our past experience with Algorithm 4 which always performed well for reasonably sized interactions. Indeed, there is a theoretical explanation to this observation. As σ goes to zero, since we chose E as the block-diagonal part of D, E and D tend to be identical. Moreover, since in our example, K and J are also identical, it is obvious that by picking $\varepsilon = \rho = 1$ in (20), Algorithm 4 is immediately stationary, which means that τ tends to zero with σ. A similar statement *cannot* be made for either Algorithm 6 or its sequential or implicit versions. Therefore, the fixed point coordination strategy should be preferred for weakly coupled subproblems. However, in our example, the *implicit* Arrow-Hurwicz coordination strategy is already better for $\sigma = 0.5$ (this is not the case with the other versions).

Let us now come back to the comparison of the two sequential versions (Algorithms 7 and 8). Since τ can be properly defined only when the recurrence is linear, that is, when K and J are quadratic, we limit ourselves to this case. We come back to the notations used in Theorem 1 and we assume that $b = 0$ and $d = 0$ without loss of

TABLE 1.

Each entry of the table reads as

	ε	
	ρ_1	τ
	ρ_2	

σ	Algorithm 4	Algorithm 6	Algorithms 7 & 8	Algorithm 9
0.5	1.00 none 0.3536 none	0.80 1.20 0.8013 0.10	1.00 2.10 0.5462 0.50	1.00 2.50 0.2500 10.00
1.0	1.00 none 0.7071 none	0.50 1.00 0.8716 0.10	0.80 1.70 0.6468 0.40	1.00 2.10 0.4095 1.10
2.0	does not converge	0.50 1.00 0.8851 0.05	0.50 1.10 0.8139 0.20	1.00 1.00 0.5774 0.50
5.0	does not converge	0.50 1.20 0.9148 0.05	1.00 1.20 0.8216 0.05	1.00 1.20 0.7038 0.05

[12] Indeed, $(p_i)_{i=1,2}$ must also be computed at each iteration but need not be stored for the next iteration.

[13] Another advantage of the implicit version which does not appear in the table is that the value of the best τ seems to be less sensitive to those of ε, ρ_1 and ρ_2 than with the other algorithms.

generality (the optimal primal and dual solutions are thus null). The recurrence can be put into the form $My^{k+1} = Ny^k$ where y is the column vector with entries (u, p, v, q). We want to prove that τ (indeed all the eigenvalues) is (are) the same with both sequential versions. An eigenvalue λ and an eigenvector y of $M^{-1}N$ obey the equation $(\lambda M - N)y = 0$. It suffices to prove that the expressions $\det(N - \lambda M)$ are formally identical for both sequential versions. For Algorithm 7, we have

$$\det(N_1 - \lambda M_1) = \begin{pmatrix} (1-\lambda)B - \varepsilon A & -\lambda\varepsilon E^* & 0 & -\varepsilon(D^* - E^*) \\ -\lambda\varepsilon E & 0 & \varepsilon I & 0 \\ 0 & \lambda\rho_1 I & (1-\lambda)I & -\rho_1 I \\ \lambda\rho_2(D-E) & 0 & \lambda\rho_2 I & (1-\lambda)I \end{pmatrix}$$

where I is the identity operator, and for Algorithm 8 we have

$$\det(N_2 - \lambda M_2) = \begin{pmatrix} (1-\lambda)B - \varepsilon A & -\lambda\varepsilon E^* & 0 & -\varepsilon(D^* - E^*) \\ -\lambda\varepsilon E & 0 & \varepsilon I & 0 \\ 0 & \lambda\rho_1 I & (1-\lambda)I & -\lambda\rho_1 I \\ \lambda\rho_2(D-E) & 0 & \rho_2 I & (1-\lambda)I \end{pmatrix}.$$

It is easy to check that, if Q_λ denotes the matrix diag $(I, I, \varepsilon I/\lambda\rho_1, -\varepsilon/\lambda\rho_2)$, $Q_\lambda(N_2 - \lambda M_2)$ is the transpose of $Q_\lambda(N_1 - \lambda M_1)$. Hence both have the same determinant, which proves the desired result since $\det Q_\lambda$ is not identically null.

6. Conclusion. We have proposed a new coordination strategy under several versions (parallel, sequential, implicit) for such decomposition techniques as Takahara's algorithm, the IPP, or one-level algorithms (in the framework of the APP), all these decomposition methods falling into the same category. Because the convergence conditions stated for this new coordination strategy avoid some prerequisite conditions involved in the convergence proof of one-level algorithms, we claim that this new coordination scheme is of broader applicability and that it is particularly suited for subproblems having interactions of large magnitude. This fact has been confirmed by some simple numerical experiments. However, we observed that for weak interactions the fixed point strategy may be competitive. Also, this latter strategy is easier to implement since only one parameter ε has to be tuned instead of three for the new scheme. Some hints are given at the end of the convergence proof in the following appendix concerning the choice of the latter.

The new coordination strategy is indeed an Arrow-Hurwicz algorithm for two auxiliary variables (one primal, the other dual) artificially introduced by some manipulation of the original constrained optimization problem. We called this manipulation "the Auxiliary Constraint Principle" and we believe that the interest of this manipulation is broader than the particular use made here. This will perhaps be shown by future applications. About the Arrow-Hurwicz scheme, we completely studied the parallel version, but the numerical experiments have shown that the sequential version, when the primal variable is updated before the dual, which is more classical, is also more efficient. Another sequential version when the dual variable is first updated yields the same performance, at least in the case of linear iterations (quadratic objective functions and affine constraints). We also proposed an implicit version which seems to be the best of all. For the sequential and implicit versions, we can state convergence theorems and give proofs which are similar to those given for the parallel version. However, these theorems and proofs do not reflect the improvement brought by the sequential and implicit over the parallel version. This point can thus be considered an open problem.

About the directions of future research in this area, we may mention the following. We made assumptions ensuring the differentiability of the saddle function Λ introduced in § 3.3 so that it was natural to use step lengths ρ_1 and ρ_2 that we call "large" as opposed to "small" steps, the latter being steps which must tend to zero as those used in some subgradient algorithms (see [5] for example). Also ε was a "large" step because J and Θ were assumed to be differentiable. The other assumptions that contribute to the differentiability of Λ are the strong convexity of J, the assumption that all the constraints considered are affine *equality* constraints (excluding in particular the presence of a feasible subset U^f which would not be equal to the whole space $\mathcal{U}$ as in (16)) and finally the assumption that E is onto. As a matter of fact, these assumptions all together guarantee uniqueness of $\hat{u}$ and $\hat{p}$ introduced in Theorem 3. Therefore, by relaxing one of these assumptions or another, either $\hat{u}$ or $\hat{p}$ (or both) would not be necessarily unique, meaning that Λ would not be differentiable in either v or q (or both). Then, we should consider algorithms mixing "large" and "small" steps ε, ρ_1 and ρ_2 (see [5] for examples of such algorithms). Forthcoming papers will discuss some of these possibilities.

Appendix: proof of Theorem 4. Existence of a saddle point of the respective Lagrangians associated with (12) and (38) are easily derived from the assumptions. Notice in particular that (45) and the same kind of inequality for K imply that the corresponding cost functions increase at least quadratically at infinity. They also imply that u^* and u^{k+1} are uniquely defined. Uniqueness of p^{k+1} (and similarly of p^*) is derived from the assumption that E (respectively, D) is onto in the following way. Necessary and sufficient optimality conditions for (38)—respectively, (12)—are

$$K'(u^{k+1}) - K'(u^k) + \varepsilon(J'(u^k) + D^*q^k + E^*(p^{k+1} - q^k)) = 0 \tag{49}$$

$$Eu^{k+1} = v^k, \tag{50}$$

respectively,

$$J'(u^*) + D^*p^* = 0 \tag{51}$$

$$Du^* = d. \tag{52}$$

Considering for instance (49), multiply it by E. Since E is onto, EE^* has a continuous inverse. Hence we get

$$p^{k+1} = q^k - (EE^*)^{-1}E\left[\frac{1}{\varepsilon}(K'(u^{k+1}) - K'(u^k)) + J'(u^k) + D^*q^k\right] \tag{53}$$

from which we see that p^{k+1} is uniquely defined since u^{k+1} is. The same applies to p^* using (51) and the inverse of DD^* if D is assumed to be onto.

The principle of the proof consists of studying the behaviour of some "Lyapounov function" along the iterations of the algorithm, from which convergence conclusions will be reached. This Lyapounov function is defined by the following quantity (at stage k)

$$\begin{cases} \phi^k := \underbrace{(1/\varepsilon)(K(u^*) - K(u^k) - \langle K'(u^k), u^* - u^k\rangle) - (a/2)\|u^* - u^k\|^2}_{\phi_1^k} \\ \quad + \underbrace{(1/2\rho_1)\|v^* - v^k\|^2}_{\phi_2^k} + \underbrace{(1/2\rho_2)\|q^* - q^k\|^2}_{\phi_3^k} + \beta(\varepsilon/2\rho_1)\underbrace{\|\hat{p}^k - q^k\|^2}_{\phi_4^k} \end{cases} \tag{54}$$

where β is a positive constant to be chosen later and $\hat{p}^k$ (together with $\hat{u}^k$) is *uniquely* defined by the equations

$$J'(\hat{u}^k) + D^*q^k + E^*(\hat{p}^k - q^k) = 0 \tag{55}$$

$$E\hat{u}^k - v^k = 0. \tag{56}$$

Indeed $\hat{u}$ and $\hat{p}$ are those already encountered in Theorem 3 when defining Λ. Notice that ϕ_1^k is larger than $(b/\varepsilon - a)/2\|u^* - u^k\|^2$ (from (45) applied to K) and this quantity is nonnegative after the condition imposed on ε and the fact that $a \leqq A$. Therefore $\phi^k \geqq 0$.

Let us now calculate the variations $\phi_i^{k+1} - \phi_i^k$ for $i = 1, \cdots, 4$.

$$\phi_1^{k+1} - \phi_1^k = \underbrace{(1/\varepsilon)(K(u^k) - K(u^{k+1}) - \langle K'(u^k), u^k - u^{k+1}\rangle)}_{a_1}$$
$$+ \underbrace{(1/\varepsilon)(\langle K'(u^k) - K'(u^{k+1}), u^* - u^{k+1}\rangle)}_{a_2}$$
$$- \frac{a}{2}(\|u^{k+1} - u^*\|^2 - \|u^k - u^*\|^2).$$

Then $a_1 \leqq -(b/2\varepsilon)\|u^{k+1} - u^k\|^2$ and from (49) and (51)

$$a_2 = \underbrace{\langle J'(u^k) - J'(u^*), u^* - u^{k+1}\rangle}_{b_1}$$
$$+ \underbrace{\langle D^*(q^k - q^*), u^* - u^{k+1}\rangle}_{b_2} + \underbrace{\langle E^*(p^{k+1} - q^k), u^* - u^{k+1}\rangle}_{b_3}.$$

By repeated uses of (45) and (47), we get that

$$b_1 \leqq \frac{A}{2}\|u^{k+1} - u^k\|^2 - \frac{a}{2}(\|u^k - u^*\|^2 + \|u^{k+1} - u^*\|^2).$$

From (33) we get that

$$\phi_2^{k+1} - \phi_2^k = -\langle q^k - p^{k+1}, v^k - v^*\rangle + \frac{1}{2\rho_1}\|v^{k+1} - v^k\|^2$$

and we notice that the first term in the right-hand side is nothing but $-b_3$. Similarly, from (34) we derive that

$$\phi_3^{k+1} - \phi_3^k = \langle q^k - q^*, D(u^{k+1} - u^*)\rangle + \frac{\rho_2}{2}\|D(u^{k+1} - u^*)\|^2$$

and the first term in the right-hand side is equal to $-b_2$. Summarizing the calculations made so far, and setting $d := \|D\|$, we have that

$$\sum_{i=1}^{3}(\phi_i^{k+1} - \phi_i^k) \leqq \left(\frac{A}{2} - \frac{b}{2\varepsilon}\right)\|u^{k+1} - u^k\|^2 + \left(\frac{\rho_2}{2}d^2 - a\right)\|u^{k+1} - u^*\|^2$$
$$+ \underbrace{(1/2\rho_1)\|v^{k+1} - v^k\|^2}_{c}. \tag{57}$$

From (56) and setting $e := \|E\|$, we have that

$$c = \frac{1}{2\rho_1}\|E(\hat{u}^{k+1} - \hat{u}^k)\|^2 \leqq \frac{e^2}{2\rho_1}\|\hat{u}^{k+1} - \hat{u}^k\|^2. \tag{58}$$

Before considering the last variation $\phi_4^{k+1} - \phi_4^k$, let us establish some inequalities that will prove useful. For the sake of brevity, let $r^k := \hat{p}^k - q^k$ (notice that $r^k = -\Lambda_v'(v^k, q^k)$ from Theorem 3). Multiplying (55) by E, we get that

$$r^k = -(EE^*)^{-1}E(J'(\hat{u}^k) + D^*q^k). \tag{59}$$

We set $\omega := \|(EE^*)^{-1}E\|$. Then

$$\begin{aligned}\|r^{k+1}-r^k\| &\leqq \omega(A\|\hat{u}^{k+1}-\hat{u}^k\|+d\|q^{k+1}-q^k\|)\\ &\leqq \omega(A\|\hat{u}^{k+1}-\hat{u}^k\|+\rho_2 d^2\|u^{k+1}-u^*\|)\end{aligned} \tag{60}$$

using (34).

Multiplying the difference of (55) expressed for index $k+1$ and for index k by $\hat{u}^k-\hat{u}^{k+1}$ yields

$$\begin{aligned}\langle r^{k+1}-r^k, v^{k+1}-v^k\rangle &= \langle E^*(r^{k+1}-r^k), \hat{u}^{k+1}-\hat{u}^k\rangle\\ &= \langle J'(\hat{u}^{k+1})-J'(\hat{u}^k)+D^*(q^{k+1}-q^k), \hat{u}^k-\hat{u}^{k+1}\rangle\\ &\leqq -a\|\hat{u}^{k+1}-\hat{u}^k\|^2+\rho_2 d^2\|u^{k+1}-u^*\|\cdot\|\hat{u}^{k+1}-\hat{u}^k\|\end{aligned} \tag{61}$$

using the strong monotony of J' and (34) again.

Consider the difference of (49) and of (55), the latter multiplied by ε. Multiplying this difference by $\hat{u}^k-u^{k+1}$ and noticing that $E\hat{u}^k = Eu^{k+1} = v^k$, it yields

$$\begin{aligned}\varepsilon a\|\hat{u}^k-u^k\|^2 &\leqq \varepsilon\langle J'(u^k)-J'(\hat{u}^k), u^k-\hat{u}^k\rangle+\langle K'(u^{k+1})-K'(u^k), u^{k+1}-u^k\rangle\\ &= \varepsilon\langle J'(u^k)-J'(\hat{u}^k), u^k-u^{k+1}\rangle+\langle K'(u^{k+1})-K'(u^k), \hat{u}^k-u^k\rangle\\ &\leqq (B+\varepsilon A)\|u^{k+1}-u^k\|\cdot\|\hat{u}^k-u^k\|\end{aligned}$$

from which it comes that

$$\|\hat{u}^k-u^k\| \leqq \frac{B+\varepsilon A}{\varepsilon a}\|u^{k+1}-u^k\|. \tag{62}$$

Consider now the difference of (53) and (59). We get that

$$\begin{aligned}\|p^{k+1}-\hat{p}^k\| &= \|p^{k+1}-q^k-r^k\|\\ &= \left\|-(EE^*)^{-1}E\left[\frac{1}{\varepsilon}(K'(u^{k+1})-K'(u^k))+J'(u^k)-J'(\hat{u}^k)\right]\right\|\\ &\leqq \omega\left(\frac{B}{\varepsilon}\|u^{k+1}-u^k\|+A\|\hat{u}^k-u^k\|\right)\\ &\leqq \frac{\omega M_\varepsilon}{\varepsilon}\|u^{k+1}-u^k\|\end{aligned} \tag{63}$$

where we have used (62) and we have set $M_\varepsilon := B+(A/a)(B+\varepsilon A)$.

Now, we calculate the variation

$$\begin{aligned}\phi_4^{k+1}-\phi_4^k &= \|r^{k+1}\|^2-\|r^k\|^2\\ &= \|r^{k+1}-r^k\|^2+2\underbrace{\langle r^k, r^{k+1}-r^k\rangle}_{h}.\end{aligned}$$

But

$$\begin{aligned}h &= \langle\hat{p}^k-p^{k+1}+p^{k+1}-q^k, r^{k+1}-r^k\rangle\\ &= \langle\hat{p}^k-p^{k+1}+\frac{v^{k+1}-v^k}{\rho_1}, r^{k+1}-r^k\rangle\end{aligned}$$

from (33). Now, using (60), (61) and (63), it comes that

$$\phi_4^{k+1}-\phi_4^k \leqq \omega^2(A\|\hat{u}^{k+1}-\hat{u}^k\|+\rho_2 d^2\|u^{k+1}-u^*\|)^2$$
$$+\frac{2\omega^2 M_\varepsilon}{\varepsilon}\|u^{k+1}-u^k\|(A\|\hat{u}^{k+1}-\hat{u}^k\|+\rho_2 d^2\|u^{k+1}-u^*\|)$$
$$+\frac{2}{\rho_1}(-a\|\hat{u}^{k+1}-\hat{u}^k\|^2+\rho_2 d^2\|u^{k+1}-u^*\|\cdot\|\hat{u}^{k+1}-\hat{u}^k\|).$$

Finally, adding inequality (57) to the above multiplied by $\beta(\varepsilon/2\rho_1)$, we can bound the total variation of ϕ^k (see (54)) as follows:

$$\phi^{k+1}-\phi^k \leqq -\tfrac{1}{2}(\vec{V}^k)^* \mathscr{A}(\varepsilon,\rho_1,\rho_2)\vec{V}^k \tag{64}$$

where we have set

$$(\vec{V}^k)^* := (\|u^{k+1}-u^k\|, \|\hat{u}^{k+1}-\hat{u}^k\|, \|u^{k+1}-u^*\|)$$

and

$$\mathscr{A}=\begin{pmatrix} b/\varepsilon-A & -\beta\omega^2 A M_\varepsilon/\rho_1 & -\rho_2\beta\omega^2 d^2 M_\varepsilon/\rho_1 \\ -\beta\omega^2 A M_\varepsilon/\rho_1 & 2\varepsilon\beta a/\rho_1^2-(e^2+\varepsilon\beta\omega^2A^2)/\rho_1 & -\varepsilon\rho_2\beta d^2(1+\rho_1\omega^2 A)/\rho_1^2 \\ -\rho_2\beta\omega^2 d^2 M_\varepsilon/\rho_1 & -\varepsilon\rho_2\beta d^2(1+\rho_1\omega^2 A)/\rho_1^2 & 2a-\rho_2 d^2(1+\varepsilon\rho_2\beta\omega^2 d^2/\rho_1) \end{pmatrix}.$$

Suppose that we can choose ε, ρ_1 and ρ_2 in such a way that $\mathscr{A}$ is positive definite. Then, the sequence $\{\phi^k\}$ is nonincreasing. Since, as shown earlier, it is also bounded from below (by zero), it must converge and the difference of two successive terms must converge to zero. With (64), this proves that $\vec{V}^k$, and in particular $\|u^{k+1}-u^*\|$, also converge to zero. Since E is continuous, v^k converges to v^* and $v^{k+1}-v^k$ converges to zero, hence, from (33), $p^{k+1}-q^k$ also goes to zero. Therefore, a cluster point of the sequence $\{q^k\}$ (in either the strong or the weak topology), if any, is also a cluster point of $\{p^k\}$. Observe that since ϕ^k is nonincreasing, it is bounded, and thus $\{q^k\}$ is bounded and does have weak cluster points. Let $\bar{q}$ be such a cluster point. Considering (49) and passing to the limit, it is not difficult to show that the pair $(u^*, \bar{q})$ satisfies the necessary and sufficient conditions (13).

Finally, if D is onto, hence DD^* is invertible, from the difference of (49) and (51), the latter multiplied by ε, this difference being further multiplied by $(DD^*)^{-1}D$, it is easy to conclude that q^k, and thus p^k, converge to p^* (which is unique).

For the proof to be complete, it remains to show how the convergence parameters can be chosen to ensure the positive definiteness of $\mathscr{A}$. This property translates into the following three conditions:

(i) $m_1(\varepsilon) := b/\varepsilon - A > 0$

(ii) $m_2(\varepsilon,\rho_1) := [m_1(\varepsilon)(2\varepsilon\beta a-\rho_1(e^2+\varepsilon\beta\omega^2A^2))-\beta^2\omega^4A^2M_\varepsilon^2]/\rho_1^2 > 0$

(iii) $\det \mathscr{A}(\varepsilon,\rho_1,\rho_2) > 0$.

Condition (i) is equivalent to the condition imposed to ε in Theorem 4. Condition (ii) first imposes that β be chosen in the open interval $(0, 2\varepsilon m_1(\varepsilon)a/\omega^4A^2M_\varepsilon^2)$, say $\beta = \varepsilon m_1(\varepsilon)a/\omega^4A^2M_\varepsilon^2)$ to fix ideas. Then ρ_1 must belong to the open interval $(0, \bar{\rho}_1(\varepsilon))$ where

$$\bar{\rho}_1(\varepsilon)=\frac{\varepsilon(b-\varepsilon A)a^2}{(\varepsilon(b-\varepsilon A)a+e^2\omega^2M_\varepsilon^2)\omega^2A^2}.$$

Notice that $\bar{\rho}_1(\varepsilon)$ goes to zero if ε tends to either zero or b/A. Finally, once ε and ρ_1 have been fixed at admissible values, observe that $\mathscr{A}$ becomes a block-diagonal positive

definite matrix if $\rho_2 = 0$. Of course this null value is not allowed since we manipulated $1/\rho_2$ in the expression of ϕ. But, by continuity, it is clear that there exists an open interval $(0, \bar{\rho}_2(\varepsilon, \rho_1))$ in which ρ_2 may lie while $\mathscr{A}$ remains positive definite. Unfortunately, the expression of this upper bound yielded by condition (iii) is very involved. But it should be clear that this $\bar{\rho}_2$ tends to zero if either $m_1(\varepsilon)$ or $m_2(\varepsilon, \rho_1)$ does so, that is, if ε or ρ_1 approaches zero or its upper bound.

Acknowledgment. The idea of the Arrow-Hurwicz coordination strategy occurred for the first time during a conversation of the first author with Professor Philippe Mahey of the Catholic University of Rio de Janeiro, Brazil, while the latter was visiting Fontainebleau in February 1984. His contribution is gratefully acknowledged.

REFERENCES

[1] Y. TAKAHARA, *Multilevel approach to dynamic optimization*, Report SRC-50-C-64-18, Case Western Reserve University, Cleveland, Ohio, 1964.

[2] M. D. MESAROVIC, D. MACKO, AND Y. TAKAHARA, *Theory of Hierarchical Multilevel Systems*, Academic Press, New York, 1970.

[3] G. COHEN, *Optimization by decomposition and coordination: a unified approach*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 222–232.

[4] ———, *Auxiliary problem principle and decomposition of optimization problems*, J. Optim. Theory Appl., 32 (1980), pp. 277–305.

[5] G. COHEN AND D. L. ZHU, *Decomposition coordination methods in large scale optimization problems. The nondifferentiable case and the use of augmented Lagrangians*, in Advances in Large Scale Systems Theory and Applications, Vol. I, J. B. Cruz, ed., JAI Press, Greenwich, Connecticut, 1984.

[6] G. COHEN AND J. C. CULIOLI, *Decomposition and coordination in stochastic optimization*, Proc. IFAC Symp. Large Scale Systems Theory and Applications, Zurich, Switzerland, 1986.

[7] G. COHEN, *Auxiliary problem principle extended to variational inequalities*, J. Optim. Theory Appl., 59 (1988), pp. 325–333.

[8] L. S. LASDON AND J. D. SCHOEFFLER, *A multilevel technique for optimization*, Proc. JACC, Troy, New York, 1965.

[9] G. COHEN AND G. JOALLAND, *Coordination methods by the prediction principle in large dynamic constrained optimization problems*, Proc. IFAC Symp. on Large Scale Systems Theory and Applications, Udine, Italy, 1976.

[10] K. J. ARROW, L. HURWICZ, AND H. UZAWA, *Studies in Linear and Non-Linear Programming*, Stanford University Press, Stanford, CA, 1972.

[11] V. M. DANSKIN, *The Theory of Max-Min*, Springer-Verlag, Berlin, 1967.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 158-172, January 1990

008

POSITIVE PROPER EFFICIENT POINTS AND RELATED CONE RESULTS IN VECTOR OPTIMIZATION THEORY*

JERALD P. DAUER† AND RICHARD J. GALLAGHER‡

Abstract. Positive proper efficient points are defined as solutions of appropriate linear scalar optimization problems. A geometric characterization of positive proper efficient points is given as well as conditions under which the set of positive proper efficient points is dense in the set of all efficient points. It is shown that these results are applicable in the normed vector lattices $C[a, b]$, l^p, and L^p for $1 \leqq p \leqq \infty$, and that previous related results, which required the ordering cone to have a compact or weak-compact base, are not applicable in many normed vector lattices, including $C[a, b]$, l^p, and L^p for $1 \leqq p \leqq \infty$.

Key words. vector optimization, scalarization, positive proper efficient points, normed vector lattice

AMS(MOS) subject classifications. 90C31, 46A55

1. Introduction. One important problem in vector optimization theory is to identify the efficient points of a set A that are solutions of a scalar optimization problem

$$\underset{y \in A}{\text{maximize}} f(y)$$

where f is a continuous linear functional that is strictly positive on the ordering cone. We will refer to such efficient points as positive proper efficient points.

Many authors have obtained results on "proper" efficient points (e.g., see [3]-[5], [9], [14], [16], [22]). An attempt by Hurwicz [16] to give a geometric characterization of positive proper efficient points motivates the approach in § 2, where we first obtain a necessary and sufficient condition that there exists continuous linear functionals that are strictly positive on the ordering cone. We then use this condition to characterize positive proper efficient points. This characterization applies in very general settings. In particular, it does not require the space to be locally convex, and it does not require the ordering cone to have a compact or weak-compact base (assumptions used by earlier authors [4], [5], [9]). In § 3 we examine the restrictiveness of the compactness and weak-compactness assumptions on the base. In particular, we show that the positive cones (i.e., the nonnegative orthants) in the normed vector lattices $C[a, b]$, l^p, and $L^p(\mathbb{R})$, $1 \leqq p \leqq \infty$, do not have compact or weak-compact bases; and moreover, the positive cones in l^p and $L^p(\mathbb{R})$, $1 < p < \infty$, do not even have bounded bases.

A second optimization problem involves the density of the set of positive proper efficient points in the set of all efficient points. This problem has been studied by Arrow, Barankin, and Blackwell [2], Hartley [15], Borwein [5], Dauer and Saleh [9], and Jahn [17]. However, none of these density results are applicable in the normed vector lattices $C[a, b]$, l^p, and L^p, $1 \leqq p \leqq \infty$. Radner [24], in considering the problem of efficient prices for infinite horizon production systems, obtained such a density result for the space l^∞ ordered by the nonnegative orthant. In § 4 we generalize the result of Radner to obtain a density result for many normed vector lattices including the spaces $C[a, b]$, l^p, and L^p, $1 \leqq p \leqq \infty$. Section 5 is devoted to further investigating spaces and cones satisfying the hypotheses of this density theorem. In particular, it is shown that the theorem can be applied in any reflexive normed vector lattice.

* Received by the editors February 29, 1988; accepted for publication (in revised form) April 14, 1989.
† Department of Mathematics, University of Tennessee at Chattanooga, Chattanooga, Tennessee 37403.
‡ Department of Mathematical Sciences, University of Montana, Missoula, Montana 59812.

Throughout the paper the term topological vector space will mean a real Hausdorff topological vector space and will be denoted by $\mathcal{Y}$. Also, the notation for the spaces $C[a, b]$, $B[a, b]$, c, c_0, l^p, and L^p follow those used in Dunford and Schwartz [10].

2. Positive proper efficient points. Let $\mathcal{Y}$ be a (real Hausdorff) topological vector space. A subset K of $\mathcal{Y}$ is called a *cone* if $\alpha K \subseteq K$ for all $\alpha \geqq 0$. If, in addition, $K \cap (-K) = \{\theta\}$, where θ denotes the zero vector in $\mathcal{Y}$, then K is said to be a *pointed cone*. A pointed convex cone K in $\mathcal{Y}$ induces a partial ordering on $\mathcal{Y}$ by the association

$$x \leqq y \quad \text{if and only if } y - x \in K.$$

An element f in the topological dual $\mathcal{Y}^*$ of $\mathcal{Y}$ is said to be *positive* if $f(k) \geqq 0$ for all $k \in K$ and is said to be *strictly positive* if $f(k) > 0$ for all $k \in K \setminus \{\theta\}$. For convenience we use the notation

$$K^+ = \{f \in \mathcal{Y}^* : f(k) \geqq 0 \text{ for all } k \in K\},$$

$$K^{+i} = \{f \in \mathcal{Y}^* : f(k) > 0 \text{ for all } k \in K \setminus \{\theta\}\}.$$

Note that K^{+i} is not necessarily the interior of K^+. In fact, if K is the nonnegative orthant in l^p, $1 < p < \infty$, then K^+ has empty interior and yet K^{+i} is nonempty. Also note that K^{+i} may be empty. For example, if $\mathcal{Y} = B[a, b]$, the set of all bounded functions on $[a, b]$, and

$$K = \{y \in B[a, b] : y(t) \geqq 0 \text{ for all } t \in [a, b]\},$$

then K^{+i} is empty [23, p. 27]. Proposition 2.1 gives a necessary and sufficient condition for K^{+i} to be nonempty, and Remark 2.2 further discusses conditions for K^{+i} to be nonempty. We first specify some notation and remind the reader of a definition.

If A is a subset of $\mathcal{Y}$, we denote the closure of A by $\operatorname{cl}(A)$ and denote the smallest convex cone containing A by $\operatorname{cone}(A)$. If A is convex, then $\operatorname{cone}(A) = \{\lambda a : \lambda \geqq 0, a \in A\}$. A cone K is said to have a *base* B if B is convex, $\theta \notin \operatorname{cl}(B)$ and $K = \operatorname{cone}(B)$. A based cone is necessarily pointed and convex.

The following proposition characterizes cones with K^{+i} nonempty in terms of an open generator for the cone.

PROPOSITION 2.1. *Let $\mathcal{Y}$ be a topological vector space and let K be a convex cone in $\mathcal{Y}$. Then K^{+i} is nonempty if and only if there exists an open convex set U in $\mathcal{Y}$ satisfying*

(i) $\theta \notin U$; *and*

(ii) $K \subseteq \operatorname{cone}(U)$.

Proof. If K^{+i} is nonempty take any $f \in K^{+i}$ and define $U = \{y \in \mathcal{Y} : f(y) > 0\}$. Then U is an open convex set satisfying (i) and (ii).

Conversely, suppose U is an open convex set satisfying (i) and (ii). Since $\theta \notin U$ there exists $f \in \mathcal{Y}^*$ such that $f(\theta) < f(u)$ for all $u \in U$ (e.g., see [25, p. 58]). Thus $f(u) > 0$ for all $u \in U$. From (ii) it follows that $f(k) > 0$ for all $k \in K \setminus \{\theta\}$. Thus $f \in K^{+i}$. □

Remark 2.2. It is easy to see that if K^{+i} is nonempty, then K must be pointed. Moreover, if K is convex and K^{+i} is nonempty then K is based. Indeed, let $f \in K^{+i}$ and define $B = \{y \in \mathcal{Y} : f(y) = 1\} \cap K$; then B is a base for K. In locally convex spaces the converse is also true. That is, if K is a based cone, then K^{+i} is nonempty. To see this, suppose B is a base for K. Since $\theta \notin \operatorname{cl}(B)$ there exists an open convex neighborhood V of θ such that $\theta \notin B + V$. Thus $B + V$ satisfies (i) and (ii) in Proposition 2.1, and so K^{+i} is nonempty.

It should also be remarked that Krein and Rutman [21, Thm. 2.1] and Klee [20, Thm. 2.7] have shown that if $\mathcal{Y}$ is a separable normed space and K is a closed pointed convex cone then K^{+i} is nonempty.

We now turn our attention to concepts in vector optimization theory. We begin by reviewing the definition of an efficient point and defining a special class of efficient points that are the focus of this work.

DEFINITION 2.3. Let $A \subseteq \mathcal{Y}$. A point $a_0 \in A$ is said to be an *efficient* (maximal, nondominated, Pareto-optimal) *point* of A if $(A - \{a_0\}) \cap K = \{\theta\}$.

DEFINITION 2.4. Let $A \subseteq \mathcal{Y}$. A point $a_0 \in A$ is said to be a *positive proper efficient point* of A if there exists some $f \in K^{+i}$ such that $f(a_0) \geqq f(a)$ for all $a \in A$.

We will denote the efficient points of a set A by $E(A)$ and the positive proper efficient points of A by Pos (A). It is an easy exercise to show that Pos $(A) \subseteq E(A)$. Also note that if A is compact and K^{+i} is nonempty, then Pos (A) is nonempty; hence $E(A)$ is nonempty. Though this fact is easy to show, the proof that $E(A)$ is nonempty for A compact requires a more elaborate argument involving Zorn's Lemma if we do not assume K^{+i} is nonempty [6].

A few remarks concerning the notion of positive proper efficiency are in order. In both finite- and infinite-dimensional spaces many notions of "proper" efficiency have been proposed [3, p. 234], [4, p. 57], [14, p. 618], [16, p. 89], [17, p. 1003], [22, p. 488]. See [5], [9], and [27] for comparisons of the various definitions. In Hurwicz [16], Borwein [4], and Benson [3] these definitions are motivated and defined geometrically, and it is shown that Pos (A) is a subset of the set of "proper" efficient points. Moreover, the notions of proper efficiency introduced by Borwein and Benson are equivalent to positive proper efficiency when A is convex and the ordering cone has a weak-compact base [5, Thm. 1]. Since our interest is in the set Pos (A), we give the following equivalence between positive proper efficient points and certain open generators for the cone. The result is applicable in a very general setting. In particular, it does not require the ordering cone to be weak-compact based, an assumption that will be discussed in § 3.

THEOREM 2.5. *Let $\mathcal{Y}$ be a topological vector space, let K be a pointed convex cone in $\mathcal{Y}$, and let $A \subseteq \mathcal{Y}$. A point $a_0 \in A$ is a positive proper efficient point of A if and only if there exists an open convex set U in $\mathcal{Y}$ such that*

(i) $K \subseteq \text{cone}\,(U)$; *and*

(ii) $\text{cone}\,(A - \{a_0\}) \cap U = \varnothing$.

Proof. Suppose a_0 is a positive proper efficient point of A. Then there exists $f \in K^{+i}$ such that $f(a_0) \geqq f(a)$ for all $a \in A$. Let $U = \{y \in \mathcal{Y}: 1 < f(y) < 2\}$. Then U is an open convex set satisfying $K \subseteq \text{cone}\,(U)$. Now if $a \in A$, then $f(a - a_0) \leqq 0$. Thus

$$A - \{a_0\} \subseteq \{y \in \mathcal{Y}: f(y) \leqq 0\}.$$

Since the latter set is a convex cone, we have that

$$\text{cone}\,(A - \{a_0\}) \subseteq \{y \in \mathcal{Y}: f(y) \leqq 0\}.$$

Hence $\text{cone}\,(A - \{a_0\}) \cap U = \varnothing$.

Conversely, suppose U is an open convex set satisfying (i) and (ii). From (ii) there exists $f \in \mathcal{Y}^*$ and $\gamma \in \mathbb{R}$ such that

$$f(x) \leqq \gamma < f(u)$$

for all $x \in \text{cone}\,(A - \{a_0\})$ and all $u \in U$ (e.g., see [25, p. 58]). Since $\theta \in \text{cone}\,(A - \{a_0\})$ we have $\gamma \geqq 0$. Thus $f(u) > 0$ for all $u \in U$. From (i) it follows that $f \in K^{+i}$. An easy contradiction argument shows that $f(x) \leqq 0$ for all $x \in \text{cone}\,(A - \{a_0\})$. Thus $f(a) \leqq f(a_0)$ for all $a \in A$. Hence a_0 is a positive proper efficient point of A. □

3. Compact and weak-compact based cones in normed vector lattices. Many results in the theory of ordering cones and vector optimization have been proven under the

hypothesis that the cone has a compact or weak-compact base [4], [5], [9], [20]. Thus, since many commonly used normed spaces are, in fact, normed vector lattices, it is of interest to know whether a cone that arises naturally from an existent lattice structure possesses such a base. We begin our study by reviewing the concept of a normed vector lattice. The notions discussed here will be referred to again in § 5. Readers familiar with normed vector lattices may proceed to Theorem 3.1.

If $\mathcal{Y}$ is a partially ordered vector space and $A \subseteq \mathcal{Y}$, we say that $y \in \mathcal{Y}$ is the *supremum* of A, written $\sup(A)$, if y satisfies the following two conditions:

(3.1) $a \leqq y$ for all $a \in A$;

(3.2) if $z \in \mathcal{Y}$ and $a \leqq z$ for all $a \in A$, then $y \leqq z$.

Similarly, we define the *infimum* of A, written $\inf(A)$, except we replace "$\leqq$" by "$\geqq$" in (3.1) and (3.2). A partially ordered vector space $\mathcal{Y}$ is called a *vector lattice* if for every pair $x, y \in \mathcal{Y}$ both $\sup\{x, y\}$ and $\inf\{x, y\}$ exist. If $\mathcal{Y}$ is a vector lattice and if $y \in \mathcal{Y}$, we define

$$y^+ = \sup\{y, \theta\}, \quad y^- = \sup\{-y, \theta\}, \quad |y| = y^+ + y^-.$$

It is clear that y^+, y^-, and $|y|$ are in the set $K := \{y \in \mathcal{Y}: y \geqq \theta\}$, and it is easy to verify that $y = y^+ - y^-$. Thus $\mathcal{Y} = K - K$. The set K is a pointed convex cone called the *positive cone* of $\mathcal{Y}$.

If, in addition to being a vector lattice, $\mathcal{Y}$ is also a topological vector space, we say that the *lattice operations are continuous* in $\mathcal{Y}$ if the maps $y \to y^+$, $y \to y^-$ and $y \to |y|$ of $\mathcal{Y}$ into itself, and the maps $(x, y) \to \sup\{x, y\}$ and $(x, y) \to \inf\{x, y\}$ of $\mathcal{Y} \times \mathcal{Y}$ into $\mathcal{Y}$ are continuous. It can be shown that continuity of any one of the maps implies continuity of all of the maps (e.g., see [26, p. 234]).

A norm $\|\cdot\|$ on a vector lattice $\mathcal{Y}$ is said to be a *lattice norm* if $|x| \leqq |y|$ implies $\|x\| \leqq \|y\|$. A *normed vector lattice* is a vector lattice equipped with a lattice norm. It is easy to check that the lattice operations are continuous in a normed vector lattice.

Some examples of normed vector lattices are $B[a, b]$ and $C[a, b]$ each with pointwise ordering and the supremum norm, L^p for $1 \leqq p \leqq \infty$ with pointwise almost everywhere ordering and the usual norms, and l^p for $1 \leqq p \leqq \infty$ with componentwise ordering and the usual norms. The corresponding positive cones in these spaces are the nonnegative orthants.

THEOREM 3.1. *Let $\mathcal{Y}$ be a normed vector lattice and let K be the positive cone in $\mathcal{Y}$ induced by the lattice order.*

(a) *If K has a compact base, then $\mathcal{Y}$ is finite-dimensional.*

(b) *If K has a weak-compact base, then $\mathcal{Y}$ is reflexive.*

Proof. To show (a) it suffices to show that the closed unit ball is compact. To show (b) it suffices to show that the closed unit ball is weak-compact.

Let B be a compact (weak-compact) base for K. Then there exists $M > 0$ such that $\|b\| \leqq M$ for all $b \in B$ (e.g., see [25, p. 68]). Also, since $\theta \notin \operatorname{cl}(B) = B$, there exists $m > 0$ such that $\|b\| \geqq m$ for all $b \in B$. Thus, $0 < m \leqq \|b\| \leqq M$ for all $b \in B$.

Let $K_M = \{k \in K: \|k\| \leqq M\}$. Then K_M is closed and convex (and hence it is weak-closed). We claim that K_M is compact (weak-compact). To show this, first note that

$$K_M \subseteq \left\{\alpha b: 0 \leqq \alpha \leqq \frac{M}{m}, b \in B\right\}.$$

By Tychonoff's Theorem (e.g., see [10, p. 32]) $[0, M/m] \times B$ is compact in the product topology of $\mathbb{R} \times \mathcal{Y}$ where $\mathbb{R}$ has the usual topology and $\mathcal{Y}$ has the norm (weak) topology.

Also, $\mathcal{Y}$ (with either topology) is a topological vector space, and so multiplication by scalars is a continuous operation from $\mathbb{R} \times \mathcal{Y}$ into $\mathcal{Y}$. Thus, since the continuous image of a compact set is compact, the set $\{\alpha b: 0 \leqq \alpha \leqq M/m, b \in B\}$ is compact (weak-compact). Therefore, the set K_M is a closed (weak-closed) subset of a compact (weak-compact) set, and so K_M is compact (weak-compact).

Now let $\mathcal{Y}_M = \{y \in \mathcal{Y}: \|y\| \leqq M\}$. Then $\mathcal{Y}_M$ is closed (weak-closed) and $\mathcal{Y}_M \subseteq K_M - K_M$. But $K_M - K_M$ is compact (weak-compact) and hence $\mathcal{Y}_M$ is compact (weak-compact). □

COROLLARY 3.2. *The nonnegative orthants in l^1, l^∞, c, c_0, L^1, L^∞, and $C[a, b]$ are not weak-compact based.*

The converse of Theorem 3.1(a) is also true. In fact, any closed pointed convex cone in $\mathbb{R}^n$ has a compact base [20, Prop. 2.4]. However the converse of Theorem 3.1(b) is not true. Corollaries 3.4 and 3.5 show that the nonnegative orthants in the reflexive spaces l^p and $L^p(\mathbb{R})$ for $1 < p < \infty$ do not have bounded bases, and hence are not weak-compact based. We first need the following result, which is due to Johnson [19].

PROPOSITION 3.3. *Let $\mathcal{Y}$ be a normed space and let K be a convex cone in $\mathcal{Y}$. Suppose K contains a sequence $\{y_n\}$ satisfying*

(i) *there exists a constant $m > 0$ such that $\|y_n\| \geqq m$ for all n, and*

(ii) *$\{y_n\}$ converges weakly to zero.*

Then K does not have a bounded base, and hence it does not have a weak-compact base.

Proof. Suppose B is a bounded base for K. Then there exists $M > 0$ such that $\|b\| \leqq M$ for all $b \in B$. Also, for each n there exists $\alpha_n > 0$, $b_n \in B$, such that $y_n = \alpha_n b_n$. Thus $M \geqq \|b_n\| = (1/\alpha_n)\|y_n\| \geqq (1/\alpha_n)m$. Therefore $1/\alpha_n \leqq M/m$. Thus, since $\{y_n\}$ converges weakly to zero, each $f \in \mathcal{Y}^*$ satisfies

$$\lim_{n\to\infty} |f(b_n)| = \lim_{n\to\infty} \frac{1}{\alpha_n} |f(y_n)| = 0.$$

But this says that $\{b_n\}$ converges weakly to zero; that is, $\theta \in \text{wk} - \text{cl}\,(B)$. But since B is convex, $\text{wk} - \text{cl}\,(B) = \text{cl}\,(B)$ and so $\theta \in \text{cl}\,(B)$. This is a contradiction since, by definition of base, $\theta \notin \text{cl}\,(B)$. □

COROLLARY 3.4. *If $1 < p < \infty$, the nonnegative orthant in l^p does not have a bounded base, and hence it does not have a weak-compact base.*

Proof. Let $\{e_n\}$ be the standard basis. That is, $e_n = \langle \delta_{in} \rangle$ where $\delta_{in} = 0$ if $i \neq n$ and $\delta_{nn} = 1$. Then $\|e_n\| = 1$ for all n and $\{e_n\}$ converges weakly to zero. □

COROLLARY 3.5. *If $1 < p < \infty$, the nonnegative orthant in $L^p(\mathbb{R})$ with Lebesgue measure μ does not have a bounded base, and hence it does not have a weak-compact base.*

Proof. Write $\mathbb{R} = \bigcup_{n=1}^{\infty} I_n$ where $\{I_n\}$ is a disjoint collection of measurable sets such that $\mu(I_n) = 1$ for all n. Now define

$$\chi_n(x) = \begin{cases} 1 & \text{if } x \in I_n, \\ 0 & \text{if } x \notin I_n. \end{cases}$$

Then $\|\chi_n\| = 1$ for all n and $\{\chi_n\}$ converges weakly to zero. □

It should be noted that the earlier authors in vector optimization have not indicated any infinite-dimensional reflexive normed vector lattices whose positive cone is weak-compact based. However, we do have the following characterization [18, Thm. 3.8.4].

PROPOSITION 3.6. *Let $\mathcal{Y}$ be a reflexive normed vector lattice and let K be the positive cone induced by the lattice order. Then K has a weak-compact base if and only if K^+ has nonempty interior in the norm topology of $\mathcal{Y}^*$.*

4. Density of Pos (A) in E(A). In § 2 it has been noted that $\text{Pos}\,(A) \subseteq E(A)$. Many authors have given sufficient conditions for the inclusion $E(A) \subseteq \text{cl}\,[\text{Pos}\,(A)]$ to hold.

Arrow, Barankin, and Blackwell [2] seem to have been the first to obtain such a result. They assumed $\mathcal{Y} = \mathbb{R}^n$ with the ordering cone being the nonnegative orthant. Hartley [15] extended this result to arbitrary closed pointed convex ordering cones in $\mathbb{R}^n$. Radner [24] generalized the result of Arrow, Barankin, and Blackwell to the space l^∞ with the nonnegative orthant as the ordering cone. Borwein [5] and Dauer and Saleh [9] obtained density results in normed spaces that are partially ordered by a weak-compact based cone. Jahn [17] provided a density result in normed spaces partially ordered by a Bishop-Phelps cone. (A cone K is said to be a *Bishop-Phelps cone* if $K = \{y \in \mathcal{Y}: \alpha \|y\| \leqq f(y)\}$ for some $\alpha \in (0, 1]$ and some $f \in \mathcal{Y}^*$, $\|f\| = 1$.) Borwein [6, Thm. 5] stated a related density result (the set of "Borwein proper" efficient points is dense in $E(A)$) for normed spaces and required the ordering cone to be closed, convex and admit strictly positive continuous linear functionals.

In this section we give sufficient conditions for the density of Pos (A) in $E(A)$ in the setting of normed spaces without requiring the ordering cone to have a bounded base. (Note that both weak-compact based cones and Bishop-Phelps cones have bounded bases.) Instead we assume the cone satisfies property P defined below. We show that this property is satisfied by the positive cones in many normed vector lattices including $C[a, b]$, l^p, and L^p for $1 \leqq p \leqq \infty$.

For convenience we use the following notation. If $\mathcal{Y}$ is a normed space, the closed unit ball in $\mathcal{Y}^*$ is denoted by

$$\text{Ball}\,(\mathcal{Y}^*) = \{f \in \mathcal{Y}^*: \|f\| \leqq 1\}.$$

If, $g, h \in \mathcal{Y}^*$, we write $h \leqq g$ if $g - h \in K^+$.

PROPERTY P. Let $\mathcal{Y}$ be a normed space and let K be a pointed convex cone in $\mathcal{Y}$. Then K is said to satisfy property P if there exists a nonempty subset D of $K^{+i} \cap \text{Ball}\,(\mathcal{Y}^*)$ satisfying the following two conditions:

P(i) If $f, g \in D$, then there exists $h \in D$ such that $h \leqq f$ and $h \leqq g$.

P(ii) Whenever $y \in \mathcal{Y}$ and $f(y) \geqq 0$ for all $f \in D$, then $y \in K$.

The significance of Property P will become apparent in the proof of the density theorem (Theorem 4.5). First we give some examples.

Example 4.1. Let $\mathcal{Y} = \mathbb{R}^n$ and let

$$K = \{\langle y_1, y_2, \cdots, y_n\rangle: y_i \geqq 0 \text{ for } i = 1, 2, \cdots, n\}.$$

One can easily show that

$$K^{+i} = \{\langle a_1, a_2, \cdots, a_n\rangle: a_i > 0 \text{ for } i = 1, 2, \cdots, n\},$$

and that $D = K^{+i} \cap \text{Ball}\,(\mathcal{Y}^*)$ satisfies P(i) and P(ii). Hence K satisfies property P. □

Example 4.2. Let $\mathcal{Y} = l^p$, $1 \leqq p \leqq \infty$, and let

$$K = \{y := \{y_i\} \in l^p: y_i \geqq 0 \text{ for } i = 1, 2, 3, \cdots\}.$$

Choose

$$D = \{a := \{a_i\} \in l^q: \|a\|_q \leqq 1,\ a_i > 0 \text{ for } i = 1, 2, 3, \cdots\}$$

where $1/p + 1/q = 1$. Then $D \subseteq K^{+i} \cap \text{Ball}\,(\mathcal{Y}^*)$. To show that D satisfies P(i), let $a = \{a_i\}$ and $b = \{b_i\}$ be in D. For $i = 1, 2, 3, \cdots$, define $c_i = \min\{a_i, b_i\}$ and let $c = \{c_i\}$. Then $c \in D$, $c \leqq a$, and $c \leqq b$. Thus D satisfies P(i).

We now show that D satisfies P(ii). Suppose $y = \{y_i\} \in l^p$ and $y \notin K$. It suffices to show that there exists $a = \{a_i\} \in D$ such that $\sum_{i=1}^{\infty} a_i y_i < 0$. Since $y \notin K$ there exists a positive integer m such that $y_m < 0$. Also, since $l^p \subseteq l^\infty$, $\|y\|_\infty < \infty$. Thus, we define

$$a_m = \frac{\|y\|_\infty - \frac{1}{2} y_m}{\|y\|_\infty - y_m}.$$

Note that $0 < a_m < 1$. Now, let a_i, $i \neq m$, be any sequence of positive numbers such that

$$\sum_{i \neq m} a_i = 1 - a_m.$$

Let $a = \{a_i\}$. Then $a \in l^1 \subseteq l^q$ and $\|a\|_q \leqq 1$. Thus $a \in D$. Moreover,

$$\sum_{i=1}^{\infty} a_i y_i \leqq \sum_{\substack{i=1 \\ i \neq m}}^{\infty} a_i \|y\|_\infty + a_m y_m = \frac{1}{2} y_m < 0.$$

Hence, D also satisfies P(ii), and so K satisfies Property P.

Example 4.3. Let $\mathcal{Y} = L^p(\Omega, \mathcal{B}, \mu)$, $1 \leqq p \leqq \infty$, where μ is a σ-finite measure. For convenience, abbreviate $L^p(\Omega, \mathcal{B}, \mu)$ by $L^p(\mu)$. Let

$$K = \{f \in L^p(\mu): f(\omega) \geqq 0 \text{ for } \mu\text{-a.e. } \omega \in \Omega\}.$$

If $g \in L^q(\mu)$, where $1/p + 1/q = 1$, it is well known that the functional F_g, defined by

$$F_g(f) = \int_\Omega fg \, d\mu \quad \text{for all } f \in L^p(\mu), \tag{4.1}$$

is in $(L^p(\mu))^*$; and furthermore, that $\|F_g\| = \|g\|_q$. Using the notation in (4.1), define

$$D = \{F_g \in (L^p(\mu))^*: g \in L^q(\mu), \|g\| \leqq 1, g(\omega) > 0 \text{ for } \mu\text{-a.e. } \omega \in \Omega\}.$$

We show that D is a nonempty subset of $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ satisfying P(i) and P(ii).

To show that D is nonempty it suffices to show that there exists $g \in L^q(\mu)$ such that $g(\omega) > 0$ for μ-a.e. $\omega \in \Omega$. If $p = 1$ or if $\mu(\Omega) < \infty$, this is obvious; and so we assume $p \neq 1$ and $\mu(\Omega) = \infty$. Since μ is σ-finite, $\Omega = \bigcup_{i=1}^{\infty} \Omega_i$, where $0 < \mu(\Omega_i) < \infty$ for $i = 1, 2, 3, \cdots$, and $\Omega_i \cap \Omega_j = \varnothing$ for $i \neq j$. Define $g: \Omega \to \mathbb{R}$ by

$$g(\omega) = \left(\frac{1}{2^i \mu(\Omega_i)}\right)^{1/q} \quad \text{for } \omega \in \Omega_i, \quad i = 1, 2, 3, \cdots.$$

Then $g \in L^q(\mu)$ and $g(\omega) > 0$ for all $\omega \in \Omega$. Hence D is nonempty.

It is clear that $D \subseteq K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$. To show D satisfies P(i), let $F_{g_1}, F_{g_2} \in D$ and define $g \in L^q(\mu)$ by

$$g(\omega) = \min\{g_1(\omega), g_2(\omega)\} \quad \text{for all } \omega \in \Omega.$$

Then F_g, defined by (4.1), is in D, $F_g \leqq F_{g_1}$ and $F_g \leqq F_{g_2}$. Thus D satisfies P(i).

It remains to show that D satisfies P(ii). Let $f \in L^p(\mu)$, $f \notin K$. It suffices to show there exists $F_g \in K^{+i}$ such that $F_g(f) < 0$. Since $f \notin K$, the set

$$\Lambda = \{\omega \in \Omega: f(\omega) < 0\}$$

satisfies $\mu(\Lambda) > 0$. Choose $h \in L^q(\mu)$ satisfying $h(\omega) > 0$ for μ-a.e. $\omega \in \Omega$. Then $fh \in L^1(\mu)$. Let

$$\alpha = \int_\Lambda fh \, d\mu \quad \text{and} \quad \beta = \int_{\Omega \setminus \Lambda} fh \, d\mu.$$

Then $-\infty < \alpha < 0$ and $0 \leqq \beta < \infty$. Choose $\gamma > 0$ such that $\alpha + \beta\gamma < 0$, and define

$$g(\omega) = \begin{cases} h(\omega) & \text{if } \omega \in \Lambda, \\ \gamma h(\omega) & \text{if } \omega \in \Omega \setminus \Lambda. \end{cases}$$

Then F_g, defined by (4.1), is in K^{+i} and $F_g(f) < 0$. Thus, D also satisfies P(ii), and so K satisfies Property P.

Example 4.4. Let $\mathcal{Y} = C[a, b]$, the continuous functions on the interval $[a, b]$, and let

$$K = \{y \in C[a, b]: y(t) \geqq 0 \quad \text{for all } t \in [a, b]\}.$$

Let NBV$[a, b]$ denote the space of normalized functions of bounded variation on $[a, b]$. It is well known that the dual of $C[a, b]$ is isomorphic to NBV$[a, b]$. In particular, if $F \in (C[a, b])^*$, there exists a unique $v \in \text{NBV}[a, b]$ such that

$$F(y) = F_v(y) := \int_a^b y(t)\, dv(t) \quad \text{for all } y \in C[a, b]. \tag{4.2}$$

Furthermore, $\|F_v\| = \text{T.V.}(v)$, where T.V. (v) denotes the total variation of v. We denote elements in $(C[a, b])^*$ by F_v, where the subscript v is understood to be the corresponding element in NBV$[a, b]$ such that (4.2) holds.

Let

$$D = \{F_v \in (C[a, b])^*: \text{T.V.}(v) \leqq 1,\ v \text{ is continuous, piecewise linear with a finite number of linear pieces, and strictly increasing}\}.$$

Then D is a nonempty subset of $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$, and D satisfies P(i) and P(ii) [13]. Hence K satisfies Property P.

The previous examples have illustrated that the positive cones in many common normed vector lattices satisfy Property P. However, this is not always the case. Recall from § 2 that the positive cone in the normed vector lattice $B[a, b]$ does not even admit strictly positive continuous linear functionals, and hence it does not satisfy Property P. Klee [20, pp. 315–316] and Krein and Rutman [21, pp. 21–22] also give examples of pointed convex cones that do not admit strictly positive continuous linear functionals.

Below, in Theorem 4.5, the density theorem is stated. The reader should be aware that the assumption requiring the set A to be compact is more stringent than what is required on the set A by Borwein [5], Dauer and Saleh [9], and Jahn [17], who, on the other hand, place more stringent requirements on the ordering cone.

THEOREM 4.5. *Let $\mathcal{Y}$ be a normed space and let K be a pointed convex cone in $\mathcal{Y}$ satisfying Property* P. *Let A be a nonempty compact convex subset of $\mathcal{Y}$. Then* $E(A) \subseteq \text{cl}[\text{Pos}(A)]$.

Before proving this theorem it will be convenient to introduce some notation and to develop some lemmas that will be used in its proof. For the remainder of this section, let $\mathcal{Y}$ be a normed space, let K be a pointed convex cone in $\mathcal{Y}$ satisfying Property P, and let D be the subset of $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ whose existence is assumed in Property P. For each $p \in D$, define

$$K_1^+(p) = \{f \in K^+: f \geqq p \text{ and } \|f\| \leqq 1\}.$$

LEMMA 4.6. *For each $p \in D$, the set $K_1^+(p)$ is convex and weak-star compact.*

Proof. Let $p \in D$ be given. Note that

$$K_1^+(p) = (K^+ + \{p\}) \cap \text{Ball}(\mathcal{Y}^*).$$

By Alaoglu's Theorem, Ball $(\mathcal{Y}^*)$ is convex and weak-star compact. Thus, it suffices to show that K^+ is convex and weak-star closed.

Clearly, K^+ is convex. Let $\{f_\lambda\}$ be a net in K^+, and suppose that f_λ converges to f weak-star. Then $f_\lambda(y)$ converges to $f(y)$ for all $y \in \mathcal{Y}$. In particular, $f_\lambda(k)$ converges to $f(k)$ for all $k \in K$. Since $f_\lambda(k) \geqq 0$ for all λ and all $k \in K$, we get $f(k) \geqq 0$ for all $k \in K$. Hence $f \in K^+$ and so K^+ is weak-star closed. □

LEMMA 4.7. *For each $p \in D$, define the function $F: \mathcal{Y} \times K_1^+(p) \to \mathbb{R}$ by*

$$F(y,f) = f(y) \quad \text{for all } (y,f) \in \mathcal{Y} \times K_1^+(p).$$

Then F is continuous in the product topology on $\mathcal{Y} \times K_1^+(p)$, where $\mathcal{Y}$ has the norm topology and $K_1^+(p)$ has the weak-star topology.

Proof. Let $(y,f) \in \mathcal{Y} \times K_1^+(p)$ and let $\{(y_\lambda, f_\lambda)\}$ be a net in $\mathcal{Y} \times K_1^+(p)$ converging to (y,f). Then y_λ converges to y in norm and f_λ converges to f weak-star. Let $\varepsilon > 0$ be given. Let the symbol $>$ denote the ordering on the directed set for the net. Then there exists λ_1 such that $\lambda > \lambda_1$ implies that $\|y_\lambda - y\| < \varepsilon/2$, and there exists λ_2 such that $\lambda > \lambda_2$ implies $|f_\lambda(y) - f(y)| < \varepsilon/2$. Choose λ_3 such that $\lambda_3 > \lambda_1$ and $\lambda_3 > \lambda_2$. Then if $\lambda > \lambda_3$ we have

$$\begin{aligned} |F(y,f) - F(y_\lambda, f_\lambda)| &= |f(y) - f_\lambda(y_\lambda)| \\ &\leqq |f(y) - f_\lambda(y)| + |f_\lambda(y) - f_\lambda(y_\lambda)| \\ &\leqq |f(y) - f_\lambda(y)| + \|f_\lambda\| \, \|y - y_\lambda\| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

□

The following lemma is due to Fan [12, p. 121].

LEMMA 4.8. *Let A and B be compact convex sets, each in a topological vector space. Let F be a real-valued continuous function on $A \times B$. If for every fixed $b \in B$, $F(a,b)$ is a convex function of a on A, and if for every fixed $a \in A$, $F(a,b)$ is a concave function of b on B, then*

$$\min_{a \in A} \max_{b \in B} F(a,b) = \max_{b \in B} \min_{a \in A} F(a,b).$$

Consequently (e.g., see [11, p. 167]), there exists a pair $(\bar{a}, \bar{b}) \in A \times B$ satisfying

$$F(\bar{a}, b) \leqq F(\bar{a}, \bar{b}) \leqq F(a, \bar{b})$$

for all $a \in A$ and all $b \in B$.

We now have everything needed to prove Theorem 4.5.

Proof of Theorem 4.5. Since $a \in E(A)$ if and only if $\theta \in E(A - \{a\})$ and since A is compact if and only if $A - \{a\}$ is compact, we assume, without loss of generality, that $\theta \in E(A)$. We must show $\theta \in \text{cl}\,[\text{Pos}\,(A)]$.

By Lemmas 4.6 and 4.7, for each $p \in D$, the hypotheses of Lemma 4.8 are satisfied for the function $F: A \times K_1^+(p) \to \mathbb{R}$, defined by

$$F(a,f) = f(a) \quad \text{for all } (a,f) \in A \times K_1^+(p).$$

Hence, there exists $a_p \in A$, $f_p \in K_1^+(p)$ such that

$$F(a_p, f) \geqq F(a_p, f_p) \geqq F(a, f_p) \quad \text{for all } a \in A \text{ and } f \in K_1^+(p).$$

That is,

$$f(a_p) \geqq f_p(a_p) \geqq f_p(a) \quad \text{for all } a \in A \text{ and } f \in K_1^+(p). \tag{4.3}$$

Since $\theta \in A$, inequality (4.3) gives

$$f(a_p) \geqq 0 \quad \text{for all } f \in K_1^+(p). \tag{4.4}$$

Since D satisfies P(i), the pair $(D, \leqq)$ is a directed set. Hence, the set $\{a_p : p \in D\}$ is a net contained in A. Since A is compact, the net has a cluster point, say $\bar{a}$, in A. From (4.3)

$$f_p(a_p) \geqq f_p(a) \quad \text{for all } a \in A.$$

Therefore, since $f_p \in K_1^+(p) \subseteq K^{+i}$, $a_p \in \text{Pos}(A)$. Thus, the net $\{a_p : p \in D\}$ is contained in Pos (A), and so $\bar{a} \in \text{cl}[\text{Pos}(A)]$ (e.g., see [8, p.378]).

We finish the proof by showing that $\bar{a} = \theta$. Since $\theta \in E(A)$, it suffices to show that $\bar{a} \geqq \theta$. By P(ii), it suffices to show that $g(\bar{a}) \geqq 0$ for all $g \in D$. To this end, let $g \in D$ be given and let $\varepsilon > 0$ be given. It suffices to show $g(\bar{a}) > -\varepsilon$. Since $\bar{a}$ is a cluster point of $\{a_p : p \in D\}$ and since g is continuous, $g(\bar{a})$ is a cluster point of $\{g(a_p): p \in D\}$. Thus, there exists $r \in D$, $r \leqq g$, such that

$$|g(\bar{a}) - g(a_r)| < \varepsilon.$$

In particular,

$$g(\bar{a}) > g(a_r) - \varepsilon.$$

Since $r \leqq g$ we have $g \in K_1^+(r)$. Thus, by (4.4), $g(a_r) \geqq 0$. Hence $g(\bar{a}) > -\varepsilon$. □

5. Results concerning Property P. In § 4 we considered some specific examples of spaces and cones satisfying Property P. In this section we give sufficient conditions for a cone to satisfy Property P. We begin with the following proposition.

PROPOSITION 5.1. *Let $\mathcal{Y}$ be a normed vector space and let K be a pointed convex cone in $\mathcal{Y}$ satisfying Property* P. *Let X be a nontrivial subspace of $\mathcal{Y}$ and put $C = K \cap X$. Then C is a pointed convex cone in X that satisfies Property* P *(with respect to X).*

Proof. The fact that C is a pointed convex cone is clear. Let D be the subset of $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$, which satisfies P(i) and P(ii). Define

$$D_X = \{\tilde{f} : \tilde{f} = f|_X \text{ for some } f \in D\}$$

where $f|_X$ denotes the restriction of f to X. Then D_X is a nonempty subset of $C^{+i} \cap \text{Ball}(X^*)$, satisfying P(i) and P(ii) with respect to X. Hence C satisfies Property P. □

As an example of how Proposition 5.1 can be used, note that the space c of all convergent sequences and the space c_0 of all sequences converging to zero are both subspaces of l^∞. Hence, by Example 4.2, the nonnegative orthants in c and c_0 satisfy Property P.

In many of the examples given in § 4 we were able to choose the subset D of $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ satisfying P(i) and P(ii) to be $D = K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$. Specifically, this was done in Example 4.1; Example 4.2 for $1 \leqq p < \infty$; and Example 4.3 for $1 \leqq p < \infty$. Whether or not this can always be done, assuming of course that K^{+i} is nonempty, is an open question. Theorem 5.2 and Corollary 5.3 give sufficient conditions under which such a choice is possible.

The notation $[\theta, k]$ is used to denote the order interval

$$[\theta, k] := \{y \in \mathcal{Y} : \theta \leqq y \leqq k\}.$$

THEOREM 5.2. *Let $\mathcal{Y}$ be a normed vector lattice and let K be the positive cone induced by the lattice order. Suppose that*

(i) *K^{+i} is nonempty, and*

(ii) *$[\theta, k]$ is weak-compact for every $k \in K$.*

Then $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ satisfies P(i) *and* P(ii), *and hence K satisfies Property* P.

COROLLARY 5.3. *Let $\mathcal{Y}$ be a reflexive normed vector lattice and let K be the positive cone induced by the lattice order. If K^{+i} is nonempty, then $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ satisfies* P(i) *and* P(ii), *and hence K satisfies Property* P.

The remainder of this section is devoted to results leading to proofs of Theorem 5.2 and Corollary 5.3. Specifically, Theorem 5.2 follows from Propositions 5.7 and 5.9; and Corollary 5.3 follows from Theorem 5.2 and Proposition 5.11. We begin by giving

sufficient conditions for $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ to satisfy P(ii). The conditions are stated in Proposition 5.7. We first develop some lemmas from which the proposition will follow. The first lemma is well known; its proof uses a standard separation argument (e.g., see [16, p. 66]).

LEMMA 5.4. *Let $\mathcal{Y}$ be a locally convex space and let K be a closed convex cone in $\mathcal{Y}$. If $y \in \mathcal{Y}$ and $f(y) \geqq 0$ for all $f \in K^+$, then $y \in K$.*

LEMMA 5.5. *Let $\mathcal{Y}$ be a locally convex space and let K be a closed pointed convex cone in $\mathcal{Y}$ for which K^{+i} is nonempty. If $y \in \mathcal{Y}$ and $f(y) \geqq 0$ for all $f \in K^{+i}$, then $y \in K$.*

Proof. Suppose $y \notin K$. By Lemma 5.4 there exists $f \in K^+$ such that $f(y) < 0$. Let $g \in K^{+i}$ and choose $\alpha > 0$ such that $f(y) + \alpha g(y) < 0$. Define $F = f + \alpha g$. Then $F \in K^{+i}$ and $F(y) < 0$. □

To apply Lemma 5.5 to the proof of Theorem 5.2, the positive cone in a normed vector lattice must be shown to be closed, pointed, and convex. In § 3, it has been noted that the positive cone is pointed and convex. A proof that the positive cone is closed if the lattice operations are continuous can be found in Schaefer [26, p. 235]. For reference we record these facts in the following lemma.

LEMMA 5.6. *Let $\mathcal{Y}$ be a vector lattice and let $K = \{y \in \mathcal{Y}: y \geqq \theta\}$. Then K is convex and pointed. If, in addition to being a vector lattice, $\mathcal{Y}$ is also a topological vector space such that the lattice operations are continuous, then K is closed.*

Lemmas 5.5 and 5.6 imply the following proposition.

PROPOSITION 5.7. *Let $\mathcal{Y}$ be a normed vector lattice and let K be the positive cone induced by the lattice order. If K^{+i} is nonempty, then $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ satisfies* P(ii).

We now work toward obtaining sufficient conditions for $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ to satisfy P(i). The conditions obtained are stated in Proposition 5.9.

LEMMA 5.8. *Let $\mathcal{Y}$ be a vector lattice and let K be the positive cone induced by the lattice order. Let f and g be positive linear functionals on $\mathcal{Y}$ and define $f \wedge g: K \to \mathbb{R}$ by*

$$(f \wedge g)(k) = g(k) - \sup\{(g-f)(y): \theta \leqq y \leqq k\} \quad \text{for all } k \in K.$$

Define $h: \mathcal{Y} \to \mathbb{R}$ by

$$h(y) = (f \wedge g)(y^+) - (f \wedge g)(y^-) \quad \text{for all } y \in \mathcal{Y}.$$

Then h is a positive linear functional on $\mathcal{Y}$ such that $h \leqq f$ and $h \leqq g$.

If, in addition to being a vector lattice, $\mathcal{Y}$ is also a topological vector space such that the lattice operations are continuous, then h is continuous provided at least one of f or g is continuous.

Proof. The fact that h is a linear functional satisfying $h \leqq f$ and $h \leqq g$ can be found, for example, in Aliprantis and Burkinshaw [1, p. 189] or Schaefer [26, p. 211]. Let $k \in K$. Since $\sup\{(g-f)(y): \theta \leqq y \leqq k\} \leqq g(k)$, we have

$$\begin{aligned} h(k) &= (f \wedge g)(k) \\ &= g(k) - \sup\{(g-f)(y): \theta \leqq y \leqq k\} \\ &\geqq g(k) - g(k) \\ &= 0. \end{aligned}$$

Thus, $h(k) \geqq 0$ for all $k \in K$, and so h is positive.

Now assume that f is continuous. Let $\{y_\lambda\}$ be a net in $\mathcal{Y}$ converging to θ. Since the mapping $T: \mathcal{Y} \to \mathcal{Y}$ defined by $T(y) = |y|$ is continuous, we have that

$$|y_\lambda| = T(y_\lambda) \to T(\theta) = |\theta| = \theta.$$

Also, since $0 \leqq h(|y_\lambda|) \leqq f(|y_\lambda|)$ and since f is continuous, we get that $h(|y_\lambda|) \to 0$. That is,

$$h(y_\lambda^+) + h(y_\lambda^-) = h(y_\lambda^+ + y_\lambda^-) = h(|y_\lambda|) \to 0.$$

But $h(y_\lambda^+) \geqq 0$ and $h(y_\lambda^-) \geqq 0$, and so we have $h(y_\lambda^+) \to 0$ and $h(y_\lambda^-) \to 0$. It follows that $h(y_\lambda) \to 0$, and so h is continuous at θ. Since h is linear, h is continuous everywhere. □

The following proposition uses Lemma 5.8 to obtain sufficient conditions for $K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ to satisfy P(i).

PROPOSITION 5.9. *Let $\mathcal{Y}$ be a normed vector lattice and let K be the positive cone induced by the lattice order. Suppose that*

(i) *K^{+i} is nonempty, and*

(ii) *$[\theta, k]$ is weak-compact for every $k \in K$.*

Then for every pair $f, g \in K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ there exists $h \in K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ such that $h \leqq f$ and $h \leqq g$.

Proof. Let $f, g \in K^{+i} \cap \text{Ball}(\mathcal{Y}^*)$ and define h as in Lemma 5.8. Then $h \in K^+$, $h \leqq f$, and $h \leqq g$. Also, since the normed dual of a normed vector lattice is a Banach lattice [26, p. 238], it follows that $\|h\| \leqq \|f\| \leqq 1$. Hence $h \in \text{Ball}(\mathcal{Y}^*)$.

It remains to show that $h \in K^{+i}$; that is, to show $h(k) > 0$ for all $k \in K \setminus \{\theta\}$. Let $k \in K \setminus \{\theta\}$. Then

$$\begin{aligned} h(k) &= (f \wedge g)(k) \\ &= g(k) - \sup\{(g-f)(y) : \theta \leqq y \leqq k\}. \end{aligned}$$

If $h(k) = 0$, then for every positive integer n there exists $y_n \in [\theta, k]$ such that

$$0 \leqq g(k) - (g-f)(y_n) < \frac{1}{n}.$$

That is,

$$0 \leqq g(k) - g(y_n) + f(y_n) < \frac{1}{n}.$$

Since both f and g are in K^{+i} we have

$$0 \leqq g(k) - g(y_n) < \frac{1}{n} \tag{5.1}$$

and

$$0 \leqq f(y_n) < \frac{1}{n}. \tag{5.2}$$

Now $[\theta, k]$ is weak-compact and so contains a cluster point $\bar{y}$ of $\{y_n\}$. Thus, for every weak open set U containing $\bar{y}$ and for every positive integer N, there exists $n \geqq N$ such that $y_n \in U$.

Let N be a positive integer. Define $U_N = \{y \in \mathcal{Y} : f(\bar{y}) - 1/N < f(y)\}$. Then U_N is a weak-open set and $\bar{y} \in U_N$. Thus there exists $n \geqq N$ such that $y_n \in U_N$. Hence

$$f(\bar{y}) - \frac{1}{N} < f(y_n).$$

Now $f(\bar{y}) \geqq 0$ and $1/n \leqq 1/N$, so the above inequality together with (5.2) yields

$$0 \leqq f(\bar{y}) < \frac{2}{N}.$$

Since N is arbitrary, we get that $f(\bar{y}) = 0$. But $f \in K^{+i}$ and so $\bar{y} = \theta$.

Again let N be a positive integer. Define $G_N = \{y \in \mathcal{Y}: g(y) < 1/N\}$. Then G_N is weak-open and $\bar{y} = \theta \in G_N$. Thus, there exists $n \geqq N$ such that $y_n \in G_N$. From (5.1) we have

$$g(y_n) \leqq g(k) < \frac{1}{n} + g(y_n),$$

which implies that

$$0 \leqq g(k) < \frac{2}{N}.$$

Since N is arbitrary $g(k) = 0$. But this is a contradiction since $g \in K^{+i}$ and $k \in K \setminus \{\theta\}$. □

Note that Propositions 5.7 and 5.9 complete the proof of Theorem 5.2. To complete the proof of Corollary 5.3, it suffices to show that in a reflexive normed vector lattice, the order interval $[\theta, k]$ is weak-compact for every $k \in K$. This fact is proved in Proposition 5.11. We first need the following lemma.

LEMMA 5.10. *Let $\mathcal{Y}$ be a locally convex space and let K be a closed pointed convex cone in $\mathcal{Y}$. Then $[\theta, k]$ is weak-closed for every $k \in K$.*

Proof. Let $k \in K$ and let $\{y_\lambda\}$ be a net in $[\theta, k]$ such that $\{y_\lambda\}$ converges weakly to y. It must be shown that $y \in [\theta, k]$.

Since $\{y_\lambda\}$ converges weakly to y we have that $f(y_\lambda) \to f(y)$ for every $f \in \mathcal{Y}^*$ and, in particular, for every $f \in K^+$. But if $f \in K^+$, then $0 \leqq f(y_\lambda) \leqq f(k)$ for all λ. Hence $0 \leqq f(y) \leqq f(k)$ for all $f \in K^+$. By Lemma 5.4 it follows that $0 \leqq y \leqq k$. □

PROPOSITION 5.11. *Let $\mathcal{Y}$ be a reflexive normed space and let K be a closed pointed convex cone in $\mathcal{Y}$ such that if $k_1, k_2 \in K$ and $k_1 \leqq k_2$, then $\|k_1\| \leqq \|k_2\|$. Then $[\theta, k]$ is weak-compact for every $k \in K$.*

Proof. Since $\mathcal{Y}$ is reflexive, the unit ball $\{y \in \mathcal{Y}: \|y\| \leqq 1\}$ is weak-compact. Thus, the set $\{y \in \mathcal{Y}: \|y\| \leqq \|k\|\}$ is weak-compact for every $k \in K$. The hypotheses of the proposition imply that $[\theta, k] \subseteq \{y \in \mathcal{Y}: \|y\| \leqq \|k\|\}$ and, by Lemma 5.10, that $[\theta, k]$ is weak-closed. Hence $[\theta, k]$ is weak-compact. □

With Proposition 5.9 in mind, we give one final result, which states that $[\theta, k]$ is weak-compact provided K has a weak-compact base. Although the fact itself is interesting, its usefulness in the context of normed vector lattices is somewhat limited since, as seen in § 3, normed vector lattices with weak-compact based positive cones are not abundant. Moreover, by Theorem 3.1, a normed vector lattice with a weak-compact based positive cone is necessarily reflexive; and by Lemma 5.6, Lemma 5.10, and Proposition 5.11, it already follows that in a reflexive normed vector lattice the set $[\theta, k]$ is weak-compact for every k in the positive cone. However, the result is useful if one is interested in cones that do not arise from a lattice structure.

PROPOSITION 5.12. *Let $\mathcal{Y}$ be a normed space and let K be a cone in $\mathcal{Y}$ with a compact (weak-compact) base. Furthermore, suppose that if $k_1, k_2 \in K$ and $k_1 \leqq k_2$, then $\|k_1\| \leqq \|k_2\|$. Then $[\theta, k]$ is compact (weak-compact) for every $k \in K$.*

Proof. Let B be a compact (weak-compact) base for K and let $k \in K$ be given. Since $\theta \notin \mathrm{cl}\,(B) = B$, there exists $m > 0$ such that $\|b\| \geqq m$ for all $b \in B$. Since B is

weak-compact, it follows that K is closed. Thus, by Lemma 5.10, $[\theta, k]$ is weak-closed and hence closed. Also,

$$[\theta, k] \subseteq \{y \in \mathcal{Y}: y = \alpha b \text{ where } 0 \leqq \alpha \leqq \|k\|/m, b \in B\}.$$

Using Tychonoff's Theorem and the fact that scalar multiplication is continuous, it follows that the latter set is compact (weak-compact). Hence $[\theta, k]$ is compact (weak-compact). □

It should be remarked that the above proposition is true in more generality. In particular, it can be shown that if $\mathcal{Y}$ is a locally convex space and K has a compact (bounded) base, then $[\theta, k]$ is compact (bounded) for all $k \in K$ [13, Prop. 5.9]. Also, it is interesting to note that the converse of Proposition 5.12 is not true. Indeed, for l^p, $1 \leqq p < \infty$, the set $[\theta, k]$ is compact for all k in the nonnegative orthant [7, Ex. 2.8], [13, Prop. 5.10]. However, by Theorem 3.1(b) (for $p = 1$) and Corollary 3.4 (for $1 < p < \infty$), the nonnegative orthant in l^p is not weak-compact based.

We refer the reader to Borwein [7] for an excellent summary of many results and examples pertaining to order intervals, based cones, and Banach lattices.

REFERENCES

[1] C. D. ALIPRANTIS AND O. BURKINSHAW, *Principles of Real Analysis*, North Holland, New York, 1981.

[2] K. J. ARROW, E. W. BARANKIN, AND D. BLACKWELL, *Admissible points of convex sets*, in Contributions to the Theory of Games, Vol. II, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, 1953, pp. 87–92.

[3] H. P. BENSON, *An improved definition of proper efficiency for vector maximization with respect to cones*, J. Math. Anal. Appl., 71 (1979), pp. 232–241.

[4] J. BORWEIN, *Proper efficient points for maximization with respect to cones*, SIAM J. Control Optim., 15 (1977), pp. 57–63.

[5] ———, *The geometry of Pareto efficiency over cones*, Math. Operationsforsch. Statist. Ser. Optim., 11 (1980), pp. 235–248.

[6] ———, *On the existence of Pareto efficient points*, Math. Oper. Res., 8 (1983), pp. 64–73.

[7] ———, *Convex cones, minimality notions, and consequences*, in Recent Advances and Historical Development of Vector Optimization, J. Jahn and W. Krabs, eds., Springer-Verlag, New York, 1987, pp. 62–85.

[8] J. B. CONWAY, *A Course in Functional Analysis*, Springer-Verlag, New York, 1985.

[9] J. P. DAUER AND O. A. SALEH, *A characterization of proper minimal points as solutions of sublinear optimization problems*, Tech. Report, Department of Mathematics and Statistics, University of Nebraska, Lincoln, NE, 1987.

[10] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Part I, Interscience, New York, 1957.

[11] I. EKELAND AND R. TEMAN, *Convex Analysis and Variational Problems*, North Holland, New York, 1976.

[12] K. FAN, *Convex Sets and Their Applications*, Argonne National Laboratory, Applied Mathematics Division Summer Lectures, Argonne, IL, 1959.

[13] R. J. GALLAGHER, *Scalarization of vector optimization problems and properties of the positive cone in normed vector lattices*, Ph.D. dissertation, University of Nebraska, Lincoln, NE, 1988.

[14] A. M. GEOFFRION, *Proper efficiency and the theory of vector maximization*, J. Math. Anal. Appl., 22 (1968), pp. 618–630.

[15] R. HARTLEY, *On cone-efficiency, cone-convexity and cone-compactness*, SIAM J. Appl. Math., 34 (1978), pp. 211–222.

[16] L. HURWICZ, *Programming in linear spaces, studies in linear and nonlinear programming*, K. J. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958, pp. 38–102.

[17] J. JAHN, *A generalization of a theorem of Arrow, Barankin, and Blackwell*, SIAM J. Control Optim., 26 (1988), pp. 999–1005.

[18] G. JAMESON, *Ordered Linear Spaces*, Springer-Verlag, New York, 1970.

[19] G. W. JOHNSON, *Notes on weak-compact bases of cones in l^p*, private communication, University of Nebraska, Lincoln, NE, 1986.

[20] V. L. KLEE, *Separation properties of convex cones*, Proc. Amer. Math. Soc., 6 (1955), pp. 313-318.
[21] M. G. KREIN AND M. A. RUTMAN, *Linear Operators Leaving Invariant a Cone in a Banach Space*, American Mathematical Society, Providence, RI, 1950. (Translation Number 26.)
[22] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, in Proc. 2nd Berkeley Symp. on Mathematical Statistics and Probability, University of California Press, Berkeley, CA, 1950, pp. 481-492.
[23] A. L. PERESSINI, *Ordered Topological Vector Spaces*, Harper and Row, New York, 1967.
[24] R. RADNER, *A note on maximal points of convex sets in* l^{∞}, in Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability, University of California Press, Berkeley, CA, 1967, pp. 351-354.
[25] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
[26] H. H. SCHAEFER, *Topological Vector Spaces*, Macmillan, New York, 1966.
[27] D. J. WHITE, *Optimality and Efficiency*, John Wiley, New York, 1982.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 173-185, January 1990

009

ON A NECESSARY AND SUFFICIENT CONDITION FOR FINITE DIMENSIONALITY OF ESTIMATION ALGEBRAS*

LUEN-FAI TAM†, WING SHING WONG‡, AND STEPHEN S.-T. YAU§

Abstract. Ever since the technique of the Kalman-Bucy filter was popularized, there has been an intense interest in finding new classes of finite dimensional recursive filters. In the late seventies, the concept of the estimation algebra of a filtering system was introduced. It has proven to be an invaluable tool in the study of nonlinear filtering problems. In this paper, a simple algebraic necessary and sufficient condition is established for an estimation algebra of a special class of filtering systems to be finite-dimensional. Also presented is a rigorous proof of the Wei-Norman program which allows one to construct finite-dimensional recursive filters from finite dimensional estimation algebras.

Key words. nonlinear filters, solvable Lie algebra, estimation algebra

AMS(MOS) subject classifications. 17B30, 35K15, 60G35, 93E11

1. Introduction. The idea of using estimation algebras to construct finite-dimensional nonlinear filters was first proposed in Brockett and Clark [1] and Brockett [2]. The motivation came from the following Wei-Norman approach [3] of using Lie algebraic ideas to solve time varying linear differential equations. Consider the equation

$$\frac{d}{dt}X(t) = A(t)X(t) \equiv \sum_{i=1}^{m} a_i(t)A_iX(t), \qquad X(0) = X_0, \tag{1.0}$$

where X and A_i's are n by n matrices and a_i's are scalar-valued functions. Let $B_1, \cdots, B_l$ be a basis of the Lie algebra generated by $A_1, \cdots, A_m$. Then the Wei-Norman Theorem states that locally in t, $X(t)$ has a representation of the form,

$$X(t) = \exp(b_1(t)B_1) \cdots \exp(b_l(t)B_l)X_0, \tag{1.1}$$

where b_i's satisfy an ordinary differential equation of the form

$$\frac{db_i}{dt} = c_i(b_1, \cdots, b_l), \qquad b_i(0) = 0$$

for all i. The function c_i's in the above equation are determined by the structure constants of the Lie algebra generated by the A_i's.

The extension of Wei-Norman's approach to the nonlinear filtering problem is much more complicated. Instead of an ordinary differential equation, we have to solve the Duncan-Mortensen-Zakai (DMZ) equation, which is a stochastic partial differential equation. By working on the robust form of the DMZ equation we can reduce the complexity of the problem to that of solving a time varying partial differential equation. Working independently, Steinberg [4] applied the Wei-Norman approach to solve some partial differential equations that are roughly related to the linear filtering problem. Wong in [5] constructed some new finite-dimensional estimation algebras and used the Wei-Norman approach to synthesize finite-dimensional filters. However, the systems considered in [5] are quite specific and the question whether the Wei-Norman approach works for a general system with finite-dimensional estimation algebra remains open.

* Received by the editors July 6, 1988; accepted for publication (in revised form) April 17, 1989.
† Department of Mathematics, The Chinese University of Hong Kong, Hong Kong.
‡ Room 3M-329, AT&T Bell Laboratories, Holmdel, New Jersey 07733.
§ Department of Mathematics, University of Illinois at Chicago, Box 4348, Chicago, Illinois 60680.

In this paper we examine the properties of finite-dimensional estimation algebras and the Wei-Norman approach in detail. We consider here a class of filtering systems having the property that the drift-term f of the state evolution equation is a gradient vector field. In [6], the concept of Ω is introduced, which is defined as the matrix whose i, j-element is $(\partial f_j/\partial x_i)-(\partial f_i/\partial x_j)$. For this class of filtering systems, Ω is zero. Conversely, if $\Omega=0$, then by the Poincaré Lemma, f is a gradient vector field. So, the class of filtering systems considered here is characterized by the fact that $\Omega=0$.

Motivated by the results in Wong [6] and [7], we investigate the algebraic problem of characterizing and classifying finite-dimensional exact estimation algebras. In [6], a sufficient condition of finite dimensionality is derived for certain filtering systems. In [7], a necessary condition and some theorems of the structure of the estimation algebra are demonstrated. In this paper, we derive a simple necessary and sufficient condition for an exact estimation algebra to be finite-dimensional. As an important consequence of these algebraic results, we prove that for a system with finite-dimensional exact estimation algebras, the Wei-Norman approach always leads to finite dimensional filters. The proof will be presented in § 4. The necessary and sufficient theorem presented here also leads us to prove some classification theorems of finite-dimensional exact estimation algebras, which will be presented in a forthcoming paper.

2. Basic concepts. The filtering problem considered here is based on the following signal observation model:

$$\begin{cases} dx(t)=f(x(t))\,dt+g(x(t))\,dv(t) & x(0)=x_0, \\ dy(t)=h(x(t))\,dt+dw(t) & y(0)=0, \end{cases} \tag{2.0}$$

in which x, v, y, and w, are, respectively, $\mathbb{R}^n$, $\mathbb{R}^p$, $\mathbb{R}^m$, and $\mathbb{R}^m$ valued processes, and v and w have components which are independent, standard Brownian processes. We further assume that $n=p$, f, h are C^∞ smooth, and that g is an orthogonal matrix. We will refer to $x(t)$ as the state of the system at time t and $y(t)$ as the observation at time t.

Let $\rho(t, x)$ denote the conditional probability density of the state given the observation $\{y(s)\colon 0\leqq s\leqq t\}$. It is well known (see [8], for example) that $\rho(t, x)$ is given by normalizing a function, $\sigma(t, x)$, which satisfies the following Duncan-Mortensen-Zakai equation:

$$d\sigma(t, x)=L_0\sigma(t, x)\,dt+\sum_{i=1}^{m} L_i\sigma(t, x)\,dy_i(t), \qquad \sigma(0, x)=\sigma_0, \tag{2.1}$$

where

$$L_0=\frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2}{\partial x_i^2}-\sum_{i=1}^{n} f_i\frac{\partial}{\partial x_i}-\sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i}-\frac{1}{2}\sum_{i=1}^{m} h_i^2$$

and for $i=1, \cdots, m$, L_i is the zero degree differential operator of multiplication by h_i.[1] σ_0 is the probability density of the initial point, x_0.

Equation (2.1) is a stochastic partial differential equation. In real applications, we are interested in constructing robust state estimators from observed sample paths with some property of robustness. Davis in [9] studied this problem and proposed some robust algorithms. In our case, his basic idea reduces to defining a new unnormalized density

$$\xi(t, x)=\exp\left(-\sum_{i=1}^{m} h_i(x)y_i(t)\right)\sigma(t, x).$$

[1] If p is a vector, we use the notation p_i to represent the ith component of p.

It is easy to show that $\xi(t, x)$ satisfies the following time varying partial differential equation

$$\frac{d\xi(t,x)}{dt} = L_0\xi(t,x) + \sum_{i=1}^{m} y_i(t)[L_0, L_i]\xi(t,x) + \frac{1}{2}\sum_{i=1}^{m} y_i^2(t)[[L_0, L_i], L_i]\xi(t,x), \tag{2.2}$$

$$\xi(0, x) = \sigma_0$$

where $[\cdot, \cdot]$ is the Lie bracket as described by the following definition.

DEFINITION. If X and Y are differential operators, the Lie bracket of X and Y, $[X, Y]$, is defined by

$$[X, Y]\zeta = X(Y\zeta) - Y(X\zeta),$$

for any C^∞ function ζ.

The objective of constructing a robust finite-dimensional filter to (2.0) is equivalent to finding a smooth manifold M and complete C^∞ vector fields μ_i on M and C^∞ functions ν on $M \times \mathbb{R} \times \mathbb{R}^n$ and ω_i's on $\mathbb{R}^m$, such that $\xi(t, x)$ can be represented in the form:

$$\begin{cases} \dfrac{dz(t)}{dt} = \displaystyle\sum_{i=1}^{k} \mu_i(z(t))\omega_i(y(t)), & z(0) \in M, \qquad \text{(2.3a)} \\ \xi(t, x) = \nu(z(t), t, x). & \qquad \text{(2.3b)} \end{cases}$$

Following [10], we say that system (2.0) has a robust universal finite-dimensional filter if for each initial probability density σ_0 there exists a z_0, such that (2.3a) and (2.3b) hold if $z(0) = z_0$, and μ_i, ω_i are independent of σ_0.

In § 5, we will use the Wei-Norman approach to construct a finite-dimensional filter for (2.0). Before we can achieve that, we need to introduce the concept of the estimation algebra of (2.0) and examine its algebraic structure.

DEFINITION. The estimation algebra $\mathbf{E}$ of a filtering problem (2.0), is defined to be the Lie algebra generated by $\{L_0, L_1, \cdots, L_m\}$, or, $\mathbf{E} = \langle L_0, L_1, \cdots, L_m \rangle_{L.A.}$. If in addition there exists a potential function ϕ such that $f_i = (\partial\phi)/(\partial x_i)$ for all $1 \leqq i \leqq n$, then the estimation algebra is called exact.

From now on, unless stated otherwise, we assume the estimation algebra of (2.0) is exact. We use ∇p to denote the column vector

$$\left(\frac{\partial p}{\partial x_1}, \cdots, \frac{\partial p}{\partial x_n}\right)^T.$$

Hence, $\nabla\phi = f$.

In the case where $n = 1$, all estimation algebras are automatically exact. Note also, all exact estimation algebras are characterized by the fact that $\Omega = 0$.

Define

$$D_i = \frac{\partial}{\partial x_i} - f_i,$$

and

$$\eta = \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{m} h_i^2.$$

Then,

$$L_0 = \frac{1}{2}\left(\sum_{i=1}^{n} D_i^2 - \eta\right).$$

Recall that $f_i = (\partial\phi)/(\partial x_i)$. Hence,

$$\eta = \Delta\phi + |\nabla\phi|^2 + \sum_{i=1}^{m} h_i^2. \tag{2.4}$$

We need the following basic results for later discussion.

THEOREM 1. (Ocone). *Let* $\mathbf{E}$ *be a finite-dimensional estimation algebra. If a function* ζ *is in* $\mathbf{E}$, *then* ζ *is a polynomial of degree* $\leqq 2$.

Ocone's theorem ([11], see [12] for an extension) says that $h_1, \cdots, h_m$ in a finite-dimensional estimation algebra are polynomials of degree $\leqq 2$.

LEMMA 1. *Let* ζ *be a* C^∞ *function on* $\mathbb{R}^n$. *Suppose* $E_l(\zeta)$ *is a polynomial of degree at most k where* $E_l = \sum_{i=1}^{l} x_i\, \partial/(\partial x_i)$. *Then* $\zeta = p_k(x_1, \cdots, x_n) + \zeta(0, \cdots, 0, x_{l+1}, \cdots, x_n)$ *where* p_k *is a polynomial of degree k in* $x_1, \cdots, x_n$.

Proof.

$$\begin{aligned}
&\zeta(x_1, x_2, \cdots, x_n) - \zeta(0, \cdots, 0, x_{l+1}, \cdots, x_n) \\
&\quad = \int_0^1 \frac{d}{dt}\,\zeta(tx_1, \cdots, tx_l, x_{l+1}, \cdots, x_n)\, dt \\
&\quad = \int_0^1 \left[x_1 \frac{\partial\zeta}{\partial x_1}(tx_1, \cdots, tx_l, x_{l+1}, \cdots, x_n) + \cdots \right. \\
&\qquad\qquad \left. + x_l \frac{\partial\zeta}{\partial x_l}(tx_1, \cdots, tx_l, x_{l+1}, \cdots, x_n) \right] dt \\
&\quad = \int_0^1 E_l(\zeta)(tx_1, \cdots, tx_l, x_{l+1}, \cdots, x_n)\, dt.
\end{aligned}$$

Since $E_l(\zeta)$ is a polynomial of degree k, we see that $\int_0^1 E_l(\zeta) \times (tx_1, \cdots, tx_l, x_{l+1}, \cdots, x_n)\, dt$ is also a polynomial of degree k. □

LEMMA 2. *Let* ζ *be a* C^∞ *function on* $\mathbb{R}^n$. *Suppose* $E_l\zeta + 2\zeta$ *is a sum of polynomials of degree two and a* C^∞ *function on* $\mathbb{R}^n$ *which depends only on* $x_{l+1}, \cdots, x_n$ *variables. Then for any* $(a_{l+1}, \cdots, a_n) \in \mathbb{R}^{n-l}$, $\zeta(x_1, \cdots, x_l, a_{l+1}, \cdots, a_n)$ *is a polynomial of degree two in* $x_1, \cdots, x_l$ *variables.*

Proof. Let $\tilde{\zeta}(x_1, \cdots, x_l) = \zeta(x_1, \cdots, x_l, a_{l+1}, \cdots, a_n)$. Then

$$\begin{aligned}
E_l(\tilde{\zeta})(x_1, \cdots, x_l) + 2\tilde{\zeta}(x_1, \cdots, x_l) &= E_l(\zeta)(x_1, \cdots, x_l, a_{l+1}, \cdots, a_n) \\
&\quad + 2\zeta(x_1, \cdots, x_l, a_{l+1}, \cdots, a_n)
\end{aligned}$$

is a polynomial of degree two in $x_1, \cdots, x_l$ variables. It is well known that $\tilde{\zeta}$ can be written in the following form

$$\tilde{\zeta}(x_1, \cdots, x_l) = \text{polynomial of degree two} + \sum_{i \leqq j \leqq k} a_{ijk} x_i x_j x_k$$

where a_{ijk} are C^∞ functions on $\mathbb{R}^l$. Clearly $E_l(\tilde{\zeta}) = \text{polynomial of degree two} + \sum_{i \leqq j \leqq k} (E_l(a_{ijk}) + 3a_{ijk}) x_i x_j x_k$ and $E_l(\tilde{\zeta}) + 2\tilde{\zeta} = \text{polynomial of degree two} + \sum_{i \leqq j \leqq k} (E_l(a_{ijk}) + 5a_{ijk}) x_i x_j x_k$. This implies $\sum_{i \leqq j \leqq k} (E_l(a_{ijk}) + 5a_{ijk}) x_i x_j x_k$ is a polynomial of degree two. It follows that for each $i \leqq j \leqq k$, we have $E_l(a_{ijk}) + 5a_{ijk} = 0$. Observe that $E_l(x_1^5 a_{ijk}) = 5x_1^5 a_{ijk} + x_1^5 E_l(a_{ijk}) = 0$. In view of Lemma 1, we know that $x_1^5 a_{ijk}$ is a

polynomial of degree zero, i.e., $x_1^5 a_{ijk} = \text{constant}$. Since a_{ijk} is a C^∞ function on $\mathbb{R}^l$, we conclude that the constant is actually zero. So $\zeta(x_1, \cdots, x_l, a_{l+1}, \cdots, a_n) = \tilde{\zeta}(x_1, \cdots, x_l)$ is a polynomial of degree two in $x_1, \cdots, x_l$ variables. □

3. Structure theorems. The following theorem plays a fundamental role in the classification of exact estimation algebra. It is similar to Theorem 1 in [7], although assuming the estimation algebra is exact allows us to drop certain technical requirements on f, g and h.

THEOREM 2. *Let* **E** *be a finite-dimensional exact estimation algebra. Then* $h_1, \cdots, h_m$ *are polynomials of degree at most one.*

Proof. By Theorem 1, each h_i is a polynomial of degree at most two. Suppose h_1 is of degree two, then by using the affine transformation $\tilde{x} = Ax + b$, where A is orthogonal, we may assume h_1 is of the form

$$\sum_{i=1}^{l} c_i \tilde{x}_i^2 + \sum_{i=l+1}^{n} c_i \tilde{x}_i + c_0,$$

where $c_1, \cdots, c_l$ are nonzero real numbers, and $l \leqq n$. (If $l = n$, the second summation vanishes.) Define $\tilde{f}(\tilde{x}) = Af(x)$ and $\tilde{D}_i = \partial/\partial \tilde{x}_i - \tilde{f}_i$. If $\tilde{\phi}(\tilde{x}) = \phi(x)$, it is easy to see that

$$\tilde{f}(\tilde{x}) = \left(\frac{\partial \tilde{\phi}}{\partial \tilde{x}_1}, \cdots, \frac{\partial \tilde{\phi}}{\partial \tilde{x}_n} \right)^T.$$

Under the transformation, L_0 is mapped into:

$$\tilde{L}_0 = \frac{1}{2} \left(\sum_{i=1}^{n} \tilde{D}_i^2 - \tilde{\eta}(\tilde{x}) \right),$$

where

$$\tilde{\eta}(\tilde{x}) = \left[\sum_{i=1}^{n} \frac{\partial \tilde{f}_i(\tilde{x})}{\partial \tilde{x}_i} + \tilde{f}(\tilde{x})^T \tilde{f}(\tilde{x}) + \sum_{i=1}^{m} \tilde{h}_i(\tilde{x})^2 \right],$$

and h is transformed into

$$\tilde{h}(\tilde{x}) = h(x).$$

E is isomorphic to the Lie algebra generated by $\tilde{L}_0$ and $\tilde{h}_i$. Note that the degree of h_i in x is the same as the degree of $\tilde{h}_i$ in $\tilde{x}$. Without causing any confusion, from now on, we drop the tilde notation.

Since h_1 is not of degree one, then $l \geqq 1$. We shall produce a contradiction. Let $X_0 = h_1$, and define X_i for $i \geqq 1$ recursively by $X_i = [[L_0, X_{i-1}], X_0]$. Since $L_0 = \frac{1}{2}(\sum_{i=1}^{n} D_i^2 - \eta)$, it is easy to see that

$$X_1 = 4 \sum_{i=1}^{l} c_i^2 x_i^2 + \sum_{i=l+1}^{n} c_i^2,$$

and for $j > 1$

$$X_j = 4^j \sum_{i=1}^{l} c_i^{j+1} x_i^2.$$

By the invertibility of the Vandermonde matrix, it follows after some relabeling, if necessary, that

$$p \equiv \frac{1}{2} \sum_{i=1}^{l} x_i^2$$

is an element in **E**. Let Y_0 be the zero degree differential operator defined by multiplication by p. Define

$$Y_1=[L_0, Y_0]=\sum_{i=1}^{l} x_iD_i+l/2,$$

$$Y_2=[L_0, Y_1]=\sum_{i=1}^{l} D_i^2+\frac{1}{2}\sum_{i=1}^{l} x_i\frac{\partial\eta}{\partial x_i}=\sum_{i=1}^{l} D_i^2+\frac{1}{2}E_l(\eta),$$

and

$$Y_3=[Y_2, Y_1]=2\sum_{i=1}^{l} D_i^2-\frac{1}{2}E_l^2(\eta).$$

Then,

$$2Y_2-Y_3=\tfrac{1}{2}E_l^2(\eta)+E_l(\eta)=\tfrac{1}{2}E_l(E_l\eta+2\eta).$$

By Lemma 1, we know that $E_l(\eta)+2\eta$ is a sum of polynomial of degree two and a C^∞ function which depends on $x_{l+1}, \cdots, x_n$ variables. By Lemma 2, it follows that η is a polynomial of degree two in $x_1, \cdots, x_l$, with coefficients which are C^∞ functions in $x_{l+1}, \cdots, x_n$ only. Recall that

$$\Delta\phi+|\nabla\phi|^2=-\sum_{i=1}^{m} h_i^2+\eta. \tag{3.0}$$

Let $\psi\in C_0^\infty$ be any C^∞ function with compact support. Multiply (3.0) with ψ^2 and integrate the equation over $\mathbb{R}^n$.

$$\begin{aligned}-\int_{\mathbb{R}^n}\left(\sum_{i=1}^{m} h_i^2-\eta\right)\psi^2&=\int_{\mathbb{R}^n}\psi^2\Delta\phi+\int_{\mathbb{R}^n}\psi^2|\nabla\phi|^2\\&=-\int_{\mathbb{R}^n}2\psi\langle\nabla\psi, \nabla\phi\rangle+\int_{\mathbb{R}^n}\psi^2|\nabla\phi|^2.\end{aligned} \tag{3.1}$$

By the Schwartz inequality

$$2\int_{\mathbb{R}^n}\psi\langle\nabla\psi, \nabla\phi\rangle\leqq\int_{\mathbb{R}_n}|\nabla\psi|^2+\int_{\mathbb{R}^n}\psi^2|\nabla\phi|^2. \tag{3.2}$$

Putting (3.2) into (3.1), we get

$$\int_{\mathbb{R}^n}|\nabla\psi|^2-\int_{\mathbb{R}^n}\left(\sum_{i=1}^{m} h_i^2-\eta\right)\psi^2\geqq 0, \tag{3.3}$$

which is true for all $\psi\in C_0^\infty$. Take any nonzero C^∞ function θ with compact support. Define ψ to be θ followed by a translation in $x_1, \cdots, x_l$ variables direction. Observe that $\int_{\mathbb{R}^n}|\nabla\psi|^2$ is independent of the translation selected. On the other hand, since η is quadratic in $x_1, \cdots, x_l$ variables and h_1 is of degree four in $x_1, \cdots, x_l$, $\sum_{l=1}^{m} h_i^2-\eta$ becomes very positive when one of the $x_1, \cdots, x_l$ tends to infinity while the other variables remain fixed. We can choose translation in directions, $x_1, \cdots, x_l$, in such a way that

$$\int_{\mathbb{R}^n}\left(\sum_{i=1}^{m} h_i^2-\eta\right)\psi^2$$

is arbitrarily large while $\int_{\mathbb{R}^n} |\nabla\psi|^2$ is bounded. This of course contradicts the inequality (3.3). □

The argument above actually proves the following theorem.

THEOREM 3. *Let $F(x_1, \cdots, x_n)$ be a C^∞ function on $\mathbb{R}^n$. Suppose that there exists a path $C: \mathbb{R} \to \mathbb{R}^n$ and $\delta > 0$ such that $\lim_{t\to\infty} \|C(t)\| = \infty$ and $\lim_{t\to\infty} \sup_{B_\delta(C(t))} F = -\infty$, where $B_\delta(C(t)) = \{x \in \mathbb{R}^n \mid \|x - C(t)\| < \delta\}$. Then there is no C^∞ function ψ on $\mathbb{R}^n$ satisfying the equation*

$$\Delta\psi + |\nabla\psi|^2 = F.$$

COROLLARY. *Let $F(x_1, \cdots, x_n)$ be a polynomial on $\mathbb{R}^n$. Suppose that there exists a polynomial path $C: \mathbb{R} \to \mathbb{R}^n$ such that $\lim_{t\to\infty} \|C(t)\| = \infty$ and $\lim_{t\to\infty} F \circ C(t) = -\infty$. Then there is no C^∞ function ψ on $\mathbb{R}^n$ satisfying the equation*

$$\Delta\psi + |\nabla\psi|^2 = F.$$

Proof. It suffices to prove that $\lim_{t\to\infty} \sup_{B_\delta(C(t))} F(x_1, \cdots, x_n) = -\infty$, where $B_\delta(C(t)) = \{x \in \mathbb{R}^n \mid \|x - C(t)\| < \delta\}$ for some $\delta > 0$. Let $C(t) = (C_1(t), \cdots, C_n(t))$, where

$$\begin{aligned}
C_1(t) &= a_{11}t^k + a_{12}t^{k-1} + \cdots + a_{1k}t + b_1 \\
C_2(t) &= a_{21}t^k + a_{22}t^{k-1} + \cdots + a_{2k}t + b_2 \\
&\vdots \\
C_n(t) &= a_{n1}t^k + a_{n2}t^{k-1} + \cdots + a_{nk}t + b_n.
\end{aligned}$$

Since F is a polynomial, we have

$$(F \circ C)(t) = \gamma_1 t^d + \gamma_2 t^{d-1} + \cdots + \gamma_{d+1},$$

where $\gamma_1, \cdots, \gamma_{d+1}$ are polynomials in a_{ij} and b_i for $i \leqq n$ and $1 \leqq j \leqq k$. γ_1 must be negative since $\lim_{t\to\infty} (F \circ C)(t) = -\infty$. By continuity, we know that there exists a $\delta > 0$, and a sphere center at $(b_1, \cdots, b_n)$ with radius δ, $B_\delta(b)$, such that for any point $(b'_1, \cdots, b'_n)$ in it, the following bounds hold:

$$\begin{aligned}
&\gamma_1(a_{ij}; b'_1, \cdots, b'_n) \leqq \tfrac{1}{2}\gamma_1(a_{ij}; b_1, \cdots, b_n) < 0 \\
&|\gamma_2(a_{ij}; b'_1, \cdots, b'_n) - \gamma_2(a_{ij}; b_1, \cdots, b_n)| \leqq 1 \\
&\vdots \\
&|\gamma_{d+1}(a_{ij}; b'_1, \cdots, b'_n) - \gamma_{d+1}(a_{ij}; b_1, \cdots, b_n)| \leqq 1.
\end{aligned}$$

It follows that for $t > 0$,

$$\begin{aligned}
&\sup_{B_\delta(C(t))} F(x_1, \cdots, x_n) \\
&\quad = \sup_{b' \in B_\delta(b)} F(a_{11}t^k + \cdots + a_{1k} + b'_1, \cdots, a_{n1}t^k + \cdots + a_{nk} + b'_n) \\
&\quad = \sup_{b' \in B_\delta(b)} \{\gamma_1(a_{ij}; b'_1, \cdots, b'_n)t^d + \gamma_2(a_{ij}; b'_1, \cdots, b'_n)t^{d-1} + \cdots \\
&\qquad\qquad + \gamma_{d+1}(a_{ij}; b'_1, \cdots, b'_n)\} \\
&\quad \leqq \tfrac{1}{2}\gamma_1(a_{ij}; b_1, \cdots, b_n)t^d + (1 + \gamma_2(a_{ij}; b_1, \cdots, b_n))t^{d-1} + \cdots \\
&\qquad\qquad + (1 + \gamma_{d+1}(a_{ij}; b_1, \cdots, b_n)).
\end{aligned}$$

As $\gamma_1(a_{ij}; b_1, \cdots, b_n)$ is negative, the right-hand side tends to $-\infty$ as t tends to ∞. The assertion follows immediately. □

The following result provides a simple characterization of when the dimension of an estimation algebra is finite.

THEOREM 4. *Suppose* **E** *is an exact estimation algebra. Then,* **E** *is finite-dimensional if and only if* $\nabla h_i^T J_\eta^j$ *is a constant for* $1 \leqq i \leqq m$ *and all* $j = 0, 1, \cdots$, *where* $J_\eta = (\partial^2 \eta)/(\partial x_i\, \partial x_j)$, *denote the Hessian matrix of* η.

Proof. The sufficiency of the condition follows from the main theorem of [2]. For completeness reason, we provide the proof here. Assume the condition in Theorem 4 holds. Note that **E** is generated by $L_0, L_1, \cdots, L_m$. Recall that for $i = 1, \cdots, m$ we define

$$L_{m+i} = [L_0, L_i] = \nabla h_i^T D,$$

where D denotes the vector

$$(D_1, \cdots, D_n)^T.$$

Define **F** to be the linear space generated by first and zero degree differential operators of the form $\nabla h_i^T J_\eta^j D$ and $\nabla h_i^T J_\eta^j \nabla \eta$, for $i = 1, \cdots, m$, $j = 0, 1, \cdots$. Clearly, $L_{m+1}, \cdots, L_{2m}$ are elements in **F**. Using our stated assumption, it is also straightforward to check that

(i) $[X, Y] =$ constant if $X, Y \in \mathbf{F}$,

(ii) $[L_0, X] \in \mathbf{F}$ if $X \in \mathbf{F}$,

(iii) $[h_i, X] =$ constant for $i = 1, \cdots, m$ and X in **F**.

Conditions (i), (ii), and (iii) imply that

$$\dim \mathbf{E} \leqq \dim \operatorname{span} \{L_0, h_1, \cdots, h_m, 1\} + \dim \mathbf{F}.$$

By our stated assumption,

$$\mathbf{F} \subset \operatorname{span} \{\partial \eta/\partial x_1, \cdots, \partial \eta/\partial x_n, \partial/\partial x_1, \cdots, \partial/\partial x_n\}.$$

It follows that dimension of **E** is finite.

To prove the necessary condition, assume **E** is finite-dimensional and the condition in Theorem 4 does not hold. Without loss of generality, we may assume there is a $k \geqq 0$, such that $\nabla h_1^T, \nabla h_1^T J_\eta, \cdots, \nabla h_1^T J_\eta^k$ are constant vectors, but $\nabla h_1^T J_\eta^{k+1}$ is not. (Notice that ∇h_1 is a constant vector by Theorem 2.) Let $c = \nabla h_1$. Hence, $c^T x$, $c^T \nabla \eta$, $c^T J_\eta \nabla \eta, \cdots, c^T J_\eta^{k-1} \nabla \eta$ all have degrees at most 1. (If $k = 0$, only the first term is present.) It follows that

$$Ad_{L_0}^{2i+1} h_1 = \frac{1}{2^i} c^T J_\eta^i D \qquad i = 0, \cdots, k+1, \tag{3.4a}$$

$$Ad_{L_0}^{2i} h_1 = \frac{1}{2^i} c^T J_\eta^{i-1} \nabla \eta \qquad i = 1, \cdots, k+1. \tag{3.4b}$$

Let $b^T = c^T J_\eta^k$. There exists an orthogonal matrix, Q, such that

$$b^T Q = (d_1, 0, 0, \cdots, 0) \equiv d^T.$$

Define an orthogonal transformation on the state space by $\tilde{x} = Q^T x$. Under this new coordinate, $c^T x$ is mapped to $c^T Q \tilde{x}$, $\eta(x)$ is mapped to $\eta(Q\tilde{x})$, and $c^T J_\eta^k$ is mapped to d^T. So we may assume $b = d$. Equation (3.4) implies that D_1 and $b^T \nabla \eta = d_1(\partial \eta/\partial x_1)$ are both in **E**. By Ocone's Theorem, $(\partial \eta)/(\partial x_1)$ is a polynomial with degree at most 2. By the assumption that $\nabla h_1^T J_\eta^{k+1}$ is not a constant vector, it follows that the degree of $(\partial \eta)/(\partial x_1)$ is exactly 2. So,

$$\eta = x_1 q + r,$$

where q is a polynomial with degree 2, r is independent of x_1. Depending on the degree of q in x_1, we have three possible cases.

(i) *Degree* 2 *case.* Clearly, $\eta - \sum_{i=1}^{m} h_i^2$ can be arbitrarily negative on some polynomial path as the path tends to infinity.

(ii) *Degree* 1 *case.* It follows that $\eta = \sum_{i=2}^{n} \alpha_i x_i x_1^2 + \beta x_1 + r$, where α_i's are constants, at least one of them nonzero, β and r are independent of x_1. Clearly, $\eta - \sum_{i=1}^{m} h_i^2$ can be arbitrarily negative on some polynomial path as the path tends to infinity.

(iii) *Degree* 0 *case.* Since q is independent of x_1, $\eta = sx_1 + t$, where s and t are independent of x_1. If $\sum_{i=1}^{m} h_i^2$ is independent of x_1, then $\eta - \sum_{i=1}^{m} h_i^2$ can be arbitrarily negative. If $\sum_{i=1}^{m} h_i^2$ is dependent on x_1, it must be of degree 2 in x_1. Again, $\eta - \sum_{i=1}^{m} h_i^2$ can be arbitrarily negative on some polynomial path as the path tends to infinity.

In all three cases, there is a contradiction to the Corollary of Theorem 3. □

If $\mathbf{E}$ is finite-dimensional, then $\nabla h_i^T J_\eta^j$ is a constant for $1 \leqq i \leqq n$ and all $j = 0, 1, \cdots$. It is easy to show by inductive argument that the following theorem holds.

THEOREM 5. *Suppose* $\mathbf{E}$ *is an exact finite-dimensional estimation algebra. Then it has a basis consisting of one second degree differential operator* L_0, *first degree differential operator*(*s*) *with constant coefficients, and zero degree differential operator*(*s*) *affine in x. Moreover, if X and Y are in* $\mathbf{E}$ *with degree less than or equal to* 1, *then* $[X, Y]$ *is a constant.*

Theorem 6 follows from Theorem 5.

THEOREM 6. *An exact finite-dimensional estimation algebra is solvable.*

4. The Wei–Norman approach. In this section we will use the structural results of previous sections to derive finite-dimensional filters by the Wei-Norman-Brockett approach. To do this, the first step we have to establish is a representation analogous to (1.1).

Consider the filtering system as defined by (2.0). In the following discussion it is not necessary to assume that the estimation algebra of (2.0) is exact. However, we will retain all the notation introduced earlier. In particular, notice that (2.2) still holds. We assume that the estimation algebra is finite dimensional and has a basis consisting of $E_0 = L_0$, differential operators, $E_1, \cdots, E_p$, (for some p) of the form

$$\sum_{j=1}^{n} \alpha_{ij} D_j + \beta_i,$$

where α_{ij}'s are constants and β_i's are polynomial in x, and zero degree differential operators, $E_{p+1}, \cdots, E_q$, (for some $q > p$) affine in x. Moreover, we assume for $1 \leqq i, j \leqq p$, $[E_i, E_j]$ is a constant and that all zero degree differential operators in the estimation algebra are spanned by $E_{p+1}, \cdots, E_q$.[2]

It follows from Theorem 5 that if the estimation algebra of (2.0) is exact and finite-dimensional then it possesses such a basis. However, the exactness is not always necessary. For example, in [6] sufficient conditions are provided for nonexact systems to possess finite-dimensional estimation algebras.

It is clear that by the assumption on the basis that for $1 \leqq i, j \leqq q$,

$$[E_i, E_j] = \text{constant}.$$

For $p+1 \leqq i, j \leqq q$,

$$[E_i, E_j] = 0,$$

and for $1 \leqq i \leqq q$ the degree of $[E_0, E_i]$ as a differential operator is not greater than one.

[2] Our earlier definition of L_i still holds. Notice that the L_i's may not form a basis of the estimation algebra.

Since $[[L_0, \sum_{i=1}^{l} c_i x_i + d], \sum_{i=1}^{l} c_i x_i + d] = \sum_{i=1}^{l} c_i^2$, if c_i's and d are constants, the constant function is in the estimation algebra. Without loss of generality, we assume that E_q is the constant function 1.

For a filtering system with such a basis, $[[L_0, L_i], L_i] =$ constant for all $i = 1, \cdots, m$. Hence, $\frac{1}{2}\sum_{i=1}^{m} [[L_0, L_i], L_i] y_i^2(t)$, denoted by $u(t)$, is a function of t independent of x. Equation (2.2) becomes

$$\frac{d\xi(t,x)}{dt} = L_0\xi(t,x) + \sum_{i=1}^{m} [L_0, L_i]\xi(t,x)y_i(t) + u(t)\xi(t,x). \tag{4.0}$$

DEFINITION. Suppose X is a differential operator, ζ_0 is in the domain of X, r is a continuous function, and $R(t) = \int_0^t r(s)\,ds$. We denote by $e^{R(t)X}\zeta_0$ the solution at time t of the following equation:

$$\frac{d\zeta(t)}{dt} = r(t)X\zeta(t), \qquad \zeta(0) = \zeta_0,$$

if it is well defined.

For $1 \leqq i \leqq q$, $e^{tE_i}\zeta(x)$ can be expressed in the form $\int k(t,x,r)\zeta(r)\,dr$, for some integrable kernel k. Hence, we can extend the definition of $e^{tE_i}\zeta(x)$ to $e^{tE_i}\zeta(t,x)$, where ζ is also a function of t.

PROPOSITION 1. *If ζ is a C^∞ function in x, then for all $0 \leqq s$, the following Baker-Campbell-Hausdorff type relations hold*:

(1) *For $1 \leqq i < q$,*

$$e^{sE_i}E_0\zeta = \left(E_0 + s\sum_{i=1}^{q} a_{ij}E_j + s^2\delta_i\right)e^{sE_i}\zeta,$$

where a_{ij}'s and δ_i's are constants.

(2) *For $1 \leqq i \leqq p$, $1 \leqq j < q$, or $1 \leqq i < q$, $1 \leqq j \leqq p$,*

$$e^{sE_i}E_j\zeta = (E_j + s\gamma_{ji})e^{sE_i}\zeta,$$

where γ_{ji}'s are constants in x.

(3) *For $p+1 \leqq i, j \leqq q$, or $i = q$, $1 \leqq j \leqq q$, or $j = p$, $1 \leqq i \leqq q$,*

$$e^{sE_i}E_j\zeta = E_j e^{sE_i}\zeta.$$

Proof. If E_i is a zero degree differential operator, e^{sE_i} is simply $\exp(sE_i)$. If it is a first degree differential operator, we may assume it is of the form: $\sum_{j=1}^{n} \alpha_{ij}D_j + \beta_i$. Define α_i to be column nth-dimensional vector whose jth component is α_{ij}. Then, it is well known that

$$\begin{aligned} e^{sE_i}\zeta(x) &= \exp\left(\phi(x) - \phi(x+s\alpha_i) + \int_0^s \beta_i(x+\alpha_i(s-r))\,dr\right)\zeta(x+s\alpha_i) \\ &= \exp\left(\phi(x) - \phi(x+s\alpha_i) + \int_0^s \beta_i(x+\alpha_i r)\,dr\right)\zeta(x+s\alpha_i). \end{aligned} \tag{4.1}$$

Assume first that ϕ and ζ are analytic functions. Let $\tilde{\zeta}$ be an arbitrary analytic function in x. From our discussion, it is clear that $e^{sE_i}\tilde{\zeta}$ is well defined for all real s and $1 \leqq i \leqq q$. Moreover, for any fixed x, $e^{sE_i}E_0e^{-sE_i}\tilde{\zeta}$ is analytic in s. Hence, the classical Baker-Campbell-Hausdorff formula holds from the Taylor series expansion. That is:

$$e^{sE_i}E_0e^{-sE_i}\tilde{\zeta} = \left(E_0 + s[E_i, E_0] + \frac{s^2}{2}[E_i, [E_i, E_0]]\right)\tilde{\zeta}.$$

Now let $\tilde{\zeta} = e^{sE_i}\zeta$. By using the previously stated properties of the basis, it is easy to see that (1) holds under the analytic assumption.

Next, we relax the condition that ϕ is analytic to that it is C^∞. If E_i is a zero degree differential operator, then clearly $e^{sE_i}E_0e^{-sE_i}\tilde{\zeta}$ is still analytic in s and (1) holds as proven before. Hence, we assume that $E_i = \sum_{j=1}^n \alpha_{ij}D_j + \beta_i$. (Recall that β_i is a polynomial in x.) We can find a polynomial sequence, $\{\tilde{\phi}_i\}$, so that $\tilde{\phi}_i$ converges to ϕ and the first and second order derivatives of $\tilde{\phi}_i$ converge to the respective first and second order derivatives of ϕ. Define $\tilde{f}_{j,i}$ to be $(\partial\tilde{\phi}_j)/(\partial x_i)$ and $\tilde{D}_{j,i}$ to be $\partial/(\partial x_i) - \tilde{f}_{j,i}$. Define $\tilde{E}_{j,i}$ to be $\sum_{k=1}^n \alpha_{ik}\tilde{D}_{j,k} + \beta_i$. Finally, define

$$\tilde{E}_{j,0} = \frac{1}{2}\sum_{k=1}^n \frac{\partial^2}{\partial x_k^2} - \sum_{k=1}^n \tilde{f}_{j,k}\frac{\partial}{\partial x_k} - \sum_{k=1}^n \frac{\partial \tilde{f}_{j,k}}{\partial x_k} - \frac{1}{2}\sum_{k=1}^m h_k^2.$$

It is easy to show by (4.1) that there exist functions u and v such that:

$$\begin{aligned} &e^{s\tilde{E}_{j,i}}\tilde{E}_{j,0}e^{-s\tilde{E}_{j,i}}\tilde{\zeta} \\ &= \left(-\frac{1}{2}\sum_{k=1}^n \frac{\partial \tilde{f}_{j,k}(x)}{\partial x_k} + \frac{1}{2}\sum_{k=1}^n \tilde{f}_{j,k}^2(x) - \frac{1}{2}\sum_{k=1}^n \frac{\partial \tilde{f}_{j,k}(x+s\alpha_i)}{\partial x_k}\right. \\ &\quad \left. -\frac{1}{2}\sum_{k=1}^n \tilde{f}_{j,k}^2(x+s\alpha_i)\right)\tilde{\zeta}(x) + \sum_{k=1}^n \tilde{f}_{j,k}(x)u(s,x) + v(s,x), \end{aligned}$$

and

$$\begin{aligned} &e^{sE_i}E_0e^{-sE_i}\tilde{\zeta} \\ &= \left(-\frac{1}{2}\sum_{k=1}^n \frac{\partial f_k(x)}{\partial x_k} + \frac{1}{2}\sum_{k=1}^n f_k^2(x) - \frac{1}{2}\sum_{k=1}^n \frac{\partial f_k(x+s\alpha_i)}{\partial x_k}\right. \\ &\quad \left. -\frac{1}{2}\sum_{k=1}^n f_k^2(x+s\alpha_i)\right)\tilde{\zeta}(x) + \sum_{k=1}^n f_k(x)u(s,x) + v(s,x). \end{aligned}$$

It follows then, that

$$\lim_{j\to\infty} e^{s\tilde{E}_{j,i}}\tilde{E}_{j,0}\, e^{-s\tilde{E}_{j,i}}\tilde{\zeta} = e^{sE_i}E_0\, e^{-sE_i}\tilde{\zeta}.$$

Similarly,

$$\lim_{j\to\infty}\left(\tilde{E}_{j,0} + s[\tilde{E}_{j,i}, \tilde{E}_{j,0}] + \frac{s^2}{2}[\tilde{E}_{j,i}, [\tilde{E}_{j,i}, \tilde{E}_{j,0}]]\right)\tilde{\zeta} = \left(E_0 + s[E_i, E_0] + \frac{s^2}{2}[E_i, [E_i, E_0]]\right)\tilde{\zeta}.$$

Hence, (1) holds in this case also. For the general case, for any given x, construct sequences of analytic functions $\{\tilde{\zeta}_i\}$, so that they converge to ζ. It follows that (1) holds in the general case as well. Statements (2), (3), and (4) can be proved similarly. □

THEOREM 7. *If the estimation algebra of* (2.0) *has a basis as described earlier, then its robust* DMZ *equation* (4.0) *has a solution for all* $t \geqq 0$ *of the form*:

$$\xi(t,x) = e^{r_q(t)E_q}\cdots e^{r_1(t)E_1}\, e^{tE_0}\sigma_0, \tag{4.2}$$

where r_i's satisfy an ordinary differential equation for all t. It follows then that a universal finite-dimensional filter exists for (2.0).

Proof. Since E_0 is elliptic, for any $t>0$, $e^{tE_0}\sigma_0$ is C^∞. By differentiating $\xi(t,x)$ we have

$$\frac{d\xi(t,x)}{dt} = e^{r_qE_q}\cdots e^{r_1E_1}E_0\,e^{tE_0}\sigma_0$$

$$+\frac{dr_1}{dt}\,e^{r_qE_q}\cdots e^{r_2E_2}E_1\,e^{r_1E_1}\,e^{tE_0}\sigma_0+\cdots+\frac{dr_q}{dt}\,E_q\,e^{r_qE_q}\cdots e^{r_1E_1}\,e^{tE_0}\sigma_0.$$

By applying Proposition 1,

$$e^{r_qE_q}\cdots e^{r_1E_1}E_0\,e^{tE_0}\sigma_0 = e^{r_qE_q}\cdots e^{r_2E_2}\left(E_0+r_1\sum_{j=1}^{q}a_{1j}E_j+r_1^2\delta_1\right)e^{r_1E_1}\,e^{tE_0}\sigma_0$$

$$=\left(E_0+\sum_{i=1}^{q-1}\sum_{j=1}^{q}r_ia_{ij}E_j+\kappa_0\right)\xi(t,x),$$

where κ_0 is a polynomial in $r_1,\cdots,r_{q-1}$ and constant in x and r_q.

For $1\leqq i\leqq p$,

$$\frac{dr_i}{dt}\,e^{r_qE_q}\cdots e^{r_{i+1}E_{i+1}}E_i\,e^{r_iE_i}\,e^{r_{i-1}E_{i-1}}\cdots e^{tE_0}\sigma_0$$

$$=\frac{dr_i}{dt}\,e^{r_qE_q}\cdots e^{r_{i+2}E_{i+2}}(E_i+r_{i+1}\gamma_{i+1,i})\,e^{r_{i+1}E_{i+1}}\cdots e^{tE_0}\sigma_0$$

$$=\frac{dr_i}{dt}(E_i+\kappa_i)\xi(t,x),$$

where κ_i is a polynomial with degree 1 in r_j for $i+1\leqq j<q$ and constant in the remaining r_j's and x.

For $p+1\leqq i\leqq q$,

$$\frac{dr_i}{dt}\,e^{r_qE_q}\cdots e^{r_{i+1}E_{i+1}}E_i\,e^{r_iE_i}\,e^{r_{i-1}E_{i-1}}\cdots e^{tE_0}\sigma_0=\frac{dr_i}{dt}\,E_i\xi(t,x).$$

Hence,

$$\frac{d\xi(t,x)}{dt}=\left(E_0+\sum_{i=1}^{q-1}\sum_{j=1}^{q}r_ia_{ij}E_j+\sum_{i=1}^{q}\frac{dr_i}{dt}\,E_i+\sum_{i=1}^{p}\frac{dr_i}{dt}\,\kappa_i+\kappa_0\right)\xi(t,x). \tag{4.3}$$

By substituting (4.3) into (4.0), it is clear that ξ_t is a solution to (4.0), if for $1\leqq j<q$,

$$\frac{dr_j}{dt}=\sum_{i=1}^{m}y_i(t)e_{ij}-\sum_{i=1}^{q-1}r_ia_{ij}, \tag{4.4}$$

and

$$\frac{dr_q}{dt}=u_t+\sum_{i=1}^{m}y_i(t)e_{iq}-\sum_{i=1}^{q-1}r_ia_{iq}-\sum_{i=1}^{p}\frac{dr_i}{dt}\,\kappa_i-\kappa_0, \tag{4.5}$$

where we represent $[L_0, L_i]$ as $\sum_{j=1}^{q}e_{ij}E_j$.

By the aforementioned property of κ_i, it is clear that (4.4) and (4.5) have solutions for all t.

To see that these results lead to a finite-dimensional filter for (2.0), notice that if we let the r_i's play the role of the z_i's in (2.3a), then (4.4) and (4.5) are of the form (2.3a). By using (4.1), it is easy to check that (4.2) is of the form (2.3b). □

Remark. For the Benes systems, the β_i's are all linear. It is well known that finite-dimensional filters exist in those cases [13].

Acknowledgments. The authors thank Dr. Lawrence Ein and the reviewer for their helpful comments.

REFERENCES

[1] R. W. BROCKETT AND J. M. C. CLARK, *The geometry of the conditional density functions, in* Analysis and Optimization of Stochastic Systems, O. L. R. Jacobs et al., eds., Academic Press, New York, 1980, pp. 299-309.

[2] R. W. BROCKETT, *Nonlinear systems and nonlinear estimation theory, in* The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., Reidel, Dordrecht, 1981.

[3] J. WEI AND E. NORMAN, *On global representations of the solutions of linear differential equations as a product of exponentials*, Proc. Amer. Math. Soc., 15 (1964), pp. 327-334.

[4] S. STEINBERG, *Applications of the Lie algebraic formulas of Baker, Campbell, Hausdorff and Zassenhaus to the calculation of explicit solutions of partial differential equations*, J. Differential Equations, 26 (1979), pp. 404-434.

[5] W. S. WONG, *New classes of finite-dimensional nonlinear filters*, Systems Control Lett., 3 (1983), pp. 155-164.

[6] ———, *On a new class of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 79-83.

[7] ———, *Theorems on the structure of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 117-124.

[8] M. H. A. DAVIS AND S. I. MARCUS, *An introduction to nonlinear filtering, in* The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., Reidel, Dordrecht, 1981.

[9] M. H. A. DAVIS, *On a multiplicative functional transformation arising in nonlinear filtering theory*, Z. Wahrsch. Verw. Gebiete, 54 (1980), pp. 125-139.

[10] M. CHALEYAT-MAUREL AND D. MICHEL, *Des resultats de non existence de filtre de dimension finie*, Stochastics, 13 (1984), pp. 83-102.

[11] D. L. OCONE, *Finite dimensional estimation algebras in nonlinear filtering, in* The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., Reidel, Dordrecht, 1981.

[12] P. C. COLLINGWOOD, *Some remarks on estimation algebras*, Systems Control Lett., 7 (1986), pp. 217-224.

[13] V. BENES, *Exact finite dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1981), pp. 65-92.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 186-213, January 1990

010

OPTIMAL CONTROLS FOR STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS*

NORIAKI NAGASE† AND MAKIKO NISIO‡

Abstract. This paper deals with control problems for a state process governed by controlled linear stochastic partial differential equations, whose drift and diffusion coefficients are the second order elliptic and the first order differential operators, respectively. A relaxed system is introduced as a generalization of admissible control and the continuous dependence of state process on a relaxed system, assuming some regularity conditions, is proved. Appealing to the usual compactification method, this continuity result derives the existence of an optimal relaxed system and, under convexity condition of coefficients, an optimal relaxed system provides an optimal admissible control in a wider sense. A relaxed control can be approximated by an admissible control which is Brownian adapted and where Bellman principle holds. As an application, stochastic control of diffusions with partial observation, where the state noise and the observation noise may not be independent, is discussed.

Key words. stochastic partial differential equation, optimal control, relaxed system, weak convergence, convexity condition, Bellman principle, partial observation

AMS(MOS) subject classifications. primary 93E20; secondary 60H15

1. Introduction.

In this paper we are concerned with control problems of systems governed by the following stochastic partial differential equations (SPDE):

$$\begin{aligned} dq(t,x) = &\sum_{i,j=0}^{d} \frac{\partial}{\partial x_i}\left(a^{ij}(x, y+W(t), U(t)) \frac{\partial}{\partial x_j} q(t,x) + f^i(x, y+W(t), U(t))\right) dt \\ &+ \sum_{k=1}^{d'} \left(\sum_{i=0}^{d} b_k^i(x, y+W(t)) \frac{\partial}{\partial x_i} q(x,t) + g_k(x, y+W(t))\right) dW^k(t) \end{aligned} \tag{1.1}$$

where $W=(W^1, \cdots, W^{d'})$ is a d'-dimensional standard Wiener process and $U(t)$ an admissible control, $0 \leqq t \leqq T$, with T fixed.

The problem is to minimize a given criterion by choosing a suitable admissible control. Namely, we treat stochastic optimal controls for distributed parameter systems. The SPDE (1.1) describes intuitively a physical object governed by a partial differential equation with random perturbation, which has been investigated from various viewpoints (cf. Fujita [5], Krylov and Rozovskii [8], [9], [11], Kunita [12], Pardoux [16], Walsh [20]). But another important example is the Zakai equation for controlled partially observed diffusions (cf. [1], [2], [4], [14], [17]). In this case, inhomogeneous terms f^i and g_k are zero and b_k^i arises from the correlation between system and observation noises. Moreover, the Wiener process W is the observation process and the coefficients a^{ij} and b_k^i depend on W (cf. [4], [21]).

The main aim of this paper is to show the existence of an optimal relaxed control for systems governed by the SPDE (1.1) under the ellipticity condition (see (A.2)); in particular we assume that $(a^{ij}(x,y,u) - \frac{3}{2}\sum_{k=1}^{d'} b_k^i(x,y) b_k^j(x,y))_{i,j=1,\cdots,d}$ is nonnegative definite and some regularity conditions on the coefficients. In particular, if $b_k^i = 0$ for $i=1, \cdots, d$, $k=1, \cdots, d'$, then the matrix $(a^{ij}(x,y,u))_{i,j=1,\cdots,d}$ may be degenerate.

* Received by the editors August 10, 1988; accepted for publication (in revised form) May 1, 1989.

† Department of Mathematics and System Fundamentals, Division of System Science, Kobe University, Rokke, Kobe 657, Japan.

‡ Department of Mathematics, Faculty of Science, Kobe University, Rokko, Kobe 657, Japan.

Let Γ be a compact convex subset of $\mathbb{R}^L$. We call it a control region. Λ denotes the set of all measures on $[0, T]\times\Gamma$, such that $\lambda([0, t]\times\Gamma) = t$ for any $t\in[0, T]$. The relaxed control, which is introduced in [2] and [3], is a Λ-valued random variable (see Definition 2.1) and acts linearly on coefficients. Thus a relaxed control μ has a density μ', namely $\mu(dt, du) = \mu'(t, du)\,dt$, and when we apply a relaxed control μ, the coefficients a^{ij} and f^i are replaced by the following $\tilde{a}^{ij}$ and $\tilde{f}^i$, respectively,

$$\tilde{a}^{ij}(t, x, y+W(t), \mu) = \int_\Gamma a^{ij}(x, y+W(t), u)\mu'(t, du)$$

and

$$\tilde{f}^i(t, x, y+W(t), \mu) = \int_\Gamma f^i(x, y+W(t), u)\mu'(t, du).$$

Moreover, the system moves according to the following SPDE:

$$\begin{aligned} dq(t, x) = &\sum_{i,j=0}^{d} \frac{\partial}{\partial x_i}\left(\tilde{a}^{ij}(x, y+W(t), \mu)\frac{\partial}{\partial x_j}q(t, x) + \tilde{f}^i(x, y+W(t), \mu)\right)dt \\ &+ \sum_{k=1}^{d'}\left(\sum_{i=0}^{d} b_k^i(x, y+W(t))\frac{\partial}{\partial x_i}q(x, t) + g_k(x, y+W(t))\right)dW^k(t). \end{aligned} \tag{1.2}$$

Now Λ becomes a compact metric space, by being endowed with the weak convergence topology, and the set of all relaxed controls turns out to be a compact metric space by being endowed with the Prohorov metric. Consequently, for our purposes, it is enough to show that the solution of the SPDE (1.2) depends on the relaxed control continuously. But this is a difficult problem. We overcome this obstacle by using a method similar to that used by Nagase [14] and the evaluations for SPDE given by Krylov and Rozovskii [10]. By this means, we can prove the existence of an optimal relaxed control. Moreover, by applying our existence theorems to the Zakai equation, we can obtain an optimal control for partially observed diffusions with correlated noise (see § 7). This result is new, and is a generalization of [1]–[4] and [14].

In § 2, we will introduce several metric spaces that are appropriate to our control problems and define a relaxed system in a wider sense as a generalization of an admissible control. In § 3, we study the way in which the solution depends on the initial data and the relaxed system. In particular, we will prove the continuous dependence of the solution on the relaxed system, when we endow it with the weak convergence topology on the space of image measures of relaxed systems (Theorems 3.1 and 3.2). Section 4 is concerned with existence theorems (Theorems 4.1 and 4.2). In § 5, we will construct an approximate optimal control which is adapted to a Wiener process. Since the Wiener process in the Zakai equation is nothing but the observation process, we have an approximate optimal control, which is a function of the observed data, for partially observed diffusions. The Bellman principle will be proved in § 6 and some applications will be discussed in § 7.

2. Preliminaries. Let us define the operators L and $M = (M_1, \cdots, M_{d'})$ by

$$L(y, u)\psi(x) = \sum_{i,j=0}^{d} \partial_i(a^{ij}(x, y, u)\partial_j\psi(x) + f^i(x, y, u)) \tag{2.1}$$

and

$$M_k(y)\psi(x) = \sum_{i=0}^{d} b_k^i(x, y)\partial_i\psi(x) + g_k(x, y) \qquad \text{for } x\in\mathbb{R}^d,\ y\in\mathbb{R}^{d'},\ u\in\Gamma, \tag{2.2}$$

respectively, where $\partial_0 =$ identity and $\partial_i = \partial/\partial x_i$, $i = 1, \cdots, d$ and a^{ij}, f^i, b_k^i and g_k are bounded and uniformly continuous.

We denote by L_r^2, $r \geqq 0$, the space of real valued Borel functions on $\mathbb{R}^d$ with the norm defined by:

$$\|f\|_{0,r} = \left(\int_{\mathbb{R}^d} |(1+|x|^2)^{r/2} f(x)|^2 \, dx \right)^{1/2}.$$

Let H_r^m be the subspace of L_r^2 consisting of functions whose generalized derivatives up to the order m belong to L_r^2. Clearly H_r^m becomes a Hilbert space with the inner product

$$(f, g)_{m,r} = \sum_{|\alpha| \leqq m} \frac{|\alpha|!}{\alpha^1! \cdots \alpha^d!} \int_{\mathbb{R}^d} (1+|x|^2)^r D^\alpha f(x) D^\alpha g(x) \, dx,$$

where $\alpha = (\alpha^1, \cdots, \alpha^d)$ is a multi-index with nonnegative integer α^i, $|\alpha| = \alpha^1 + \cdots + \alpha^d$ and $D^\alpha = (\partial/\partial x_1)^{\alpha^1} \cdots (\partial/\partial x_d)^{\alpha^d}$. Let us set $\|f\|_{m,r}^2 = (f, f)_{m,r}$ and, for $r = 0$, $L_0^2 = L^2$, $H_0^m = H^m$, $(\cdot, \cdot)_{m,0} = (\cdot, \cdot)_m$ and $\|\cdot\|_{m,0} = \|\cdot\|_m$, for simplicity, if no confusion occurs.

Now we introduce the following conditions.

(A.1) $D^\alpha a^{ij}$, $D^\alpha b_k^i$ $(0 \leqq |\alpha| \leqq m+1$, $i, j = 0, 1, \cdots, d$, $k = 1, \cdots, d')$ are bounded and uniformly continuous.

(A.2) ellipticity condition: $a^{ij} = a^{ji}$, $i, j = 1, \cdots, d$, and $(a^{ij} - \frac{3}{2} b^i \cdot b^j)_{i,j=1,\cdots,d}$ is a nonnegative definite matrix, where $b^i = (b_1^i, \cdots, b_{d'}^i)$ and "$\cdot$" means the inner product in $\mathbb{R}^{d'}$.

(A.3) $f^i(\cdot, y, u)$, $g_k(\cdot, y) \in H^{m+1}$, $i = 0, \cdots, d$, $k = 1, \cdots, d'$, and their H^{m+1}-norms are bounded in $(y, u) \in \mathbb{R}^{d'} \times \Gamma$.

(A.4)$_{l,r}$ $f^i(\cdot, y, u)$, $g_k(\cdot, y) \in H_r^{l+1}$ and their H_r^{l+1}-norms are bounded in y and u.

(A.4)$_l$ For some $r > 0$, (A.4)$_{l,r}$ holds.

Hereafter we always assume (A.1) $\sim$ (A.3) and, for simplicity, we say

$$\begin{aligned} &|D^\alpha a^{ij}(x, y, u)| \leqq K, \qquad |D^\alpha b_k^i(x, y)| \leqq K, \\ &\|f^i(\cdot, y, u)\|_{m+1} \leqq K, \qquad \|g_k(\cdot, y)\|_{m+1} \leqq K. \end{aligned} \tag{2.3}$$

To study relaxed systems (in a wider sense), we need the following spaces.

By Λ we denote the set of all measures λ on $[0, T] \times \Gamma$ such that

$$\lambda([0, s] \times \Gamma) = s, \quad \text{for } s \leqq T. \tag{2.4}$$

Endowing with the weak convergence topology, we have the following proposition.

PROPOSITION 2.1. *Λ is a compact metric space (cf. [6]).*

Proof. By applying the Prohorov metric, Λ becomes a separable metric space. Suppose $\lambda_n \in \Lambda$ tends to λ weakly as $n \to \infty$. Then $\lambda_n(\cdot \times \Gamma) \to \lambda(\cdot \times \Gamma)$ weakly as a measure on $[0, T]$. Since $\lambda_n(\cdot \times \Gamma)$ is Lebesgue measure by (2.4), $\lambda(\cdot \times \Gamma)$ also satisfies (2.4). Since Λ is tight, by virtue of compactness of $[0, T] \times \Gamma$, this completes the proof. □

Let us set $\mathbf{B}(\Gamma) =$ Borel field on Γ, $\sigma_t(\Lambda) =$ the σ-field generated by $\{\lambda([0, s] \times A); s \leqq t, A \in \mathbf{B}(\Gamma)\}$ and $\sigma(\Lambda) = \sigma_T(\Lambda)$. Let $\mathcal{P} = \mathcal{P}(\Lambda)$ be the space of probabilities on $(\Lambda, \sigma(\Lambda))$, endowed with the weak convergence topology. Then Prohorov's theorem asserts the following proposition.

PROPOSITION 2.2. *$\mathcal{P}$ is a compact metric space.*

By virtue of (2.4), λ has a $\sigma_t(\Lambda)$-adapted kernel λ', namely, $\lambda(dt, du) = \lambda'(t, du)\, dt$, and $\lambda'(t, \cdot)$ is a probability on Γ for almost all t. Moreover, if λ^* is a kernel of λ, then

$\lambda'(t, \cdot) = \lambda^*(t, \cdot)$ for almost all t. Let us set

$$\tilde{h}(t, x, y, \lambda) = \int_\Gamma h(x, y, u)\lambda'(t, du) \quad \text{for } h = a^{ij} \text{ and } f^i$$

and

$$\begin{aligned} \tilde{L}(t, y, \lambda)\psi(x) &= \int_\Gamma L(y, u)\psi(x)\lambda'(t, du) \\ &= \sum_{i,j=0}^{d} \partial_i(\tilde{a}^{ij}(t, x, y, \lambda)\partial_j\psi(x) + \tilde{f}^i(t, x, y, \lambda)). \end{aligned} \tag{2.5}$$

Now we introduce a relaxed system, according to [2] and [3].

DEFINITION 2.1. $\mathcal{R} = (\Omega, \mathcal{F}, \mathcal{F}_t, P, W, \mu)$ is called a relaxed system, if

(2.6) $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ is a probability space with filteration $\mathcal{F}_t$;

(2.7) W is an $\mathcal{F}_t$-adapted d'-dimensional Wiener process with $W(0) = 0$;

and

(2.8) μ is an $\mathcal{F}_t$-adapted Λ-valued random variable (Λ-r.v. in short). Namely, $\mu(B_1 \times B_2)$ is $\mathcal{F}_t$-measurable whenever $B_1 \in \mathbf{B}[0, t]$ and $B_2 \in \sigma(\Gamma)$ (=topological σ-field on Γ).

For simplicity, we put $\mathcal{R} = (W, \mu)$, if no confusion occurs, and sometimes we call μ a relaxed control.

$\mathcal{A} = (\Omega, \mathcal{F}, \mathcal{F}_t, P, W, U)$ is an admissible system, if (2.8) is replaced by (2.9) below.

(2.9) U is a Γ-valued $\mathcal{F}_t$-adapted process.

Remark. Since $U(t)$ is regarded as $\mu'(t, \cdot) = \delta_{U(t)}$, where δ_x means δ-measure at x, $\mathcal{A}$ is also a relaxed system.

$\mathfrak{R}$ and $\mathfrak{U}$ denote the totalities of relaxed and admissible systems, respectively. Let $\pi(\mathcal{R})$ be the image measure of (W, μ) on $C(0, T; \mathbb{R}^d) \times \Lambda$. Again endowing with the weak convergence topology on the space $\Pi = \{\pi(\mathcal{R}); \mathcal{R} \in \mathfrak{R}\}$, we have the following proposition.

PROPOSITION 2.3. Π *is a compact metric space.*

Proof. The proof is easy, since W is a Wiener process and Λ is a compact metric space. □

DEFINITION 2.2. We say $\mathcal{R}_n$ converges to $\mathcal{R}$, (put $\mathcal{R}_n \to \mathcal{R}$) if, $\pi(\mathcal{R}_n) \to \pi(\mathcal{R})$ weakly.

Consider the SPDE (2.10) for $\mathcal{R} = (\Omega, \mathcal{F}, \mathcal{F}_t, P, W, \mu)$,

$$\begin{aligned} &dq(t) = \tilde{L}(t, y + W(t), \mu)q(t)\, dt + M(y + W(t))q(t)\, dW(t) \\ &q(0) = \phi \ (\in H^m). \end{aligned} \tag{2.10}$$

An H^1-valued $\mathcal{F}_t$-adapted process $q = q(\cdot, \phi, \mathcal{R})$ is called a solution of (2.10) (or a response for $\mathcal{R}$), if (2.11) and (2.12) hold.

$$E\left\{\int_0^T \|q(t)\|_{1,0}^2\, dt\right\} < \infty \tag{2.11}$$

and, for any $\eta \in C_0^\infty$ (smooth function on $\mathbb{R}^d$ with compact support) and almost all t,

$$(q(t), \eta) = (\phi, \eta) + \int_0^t \langle \tilde{L}(s, y + W(s), \mu) q(s), \eta \rangle \, ds + \int_0^t (M(y + W(s)) q(s), \eta) \, dW(s) \cdot \text{w.p. } 1 \tag{2.12}$$

holds, where $(\cdot, \cdot) = L^2(\mathbb{R}^d)$-inner product and $\langle \cdot, \cdot \rangle =$ duality pairing between H^{-1} and H^1 under $H^0 = (H^0)^*$ (= dual space of H^0), namely

$$\langle \tilde{L}(s, y + W(s), \mu) q(s), \eta \rangle = \sum_{i,j=0}^{d} (-1)^{|i|} (\tilde{a}^{ij}(s, \cdot, y + W(s), \mu) \partial_j q(s), \partial_i \eta),$$

where $|i| = 0$ (for $i = 0$), $= 1$ (for $i = 1, \cdots, d$).

Clearly (2.12) does not depend on any special choice of derivative μ'. The SPDE (2.10) can be regarded as an H^{-1}-valued SDE. (See Itô [7] for the general theory of Hilbert space valued SDE.)

According to [10] and [11], we see the following theorem.

THEOREM 2.1 (Krylov and Rozovskii). (I) *Suppose the conditions* (A.1)$\sim$(A.3). *Then, the* SPDE (2.10) *has a unique solution*

$$q \in L^2([0, T] \times \Omega; H^m) \cap L^2(\Omega: C(0, T; H^{m-1})).$$

$q(t)$ *is a Borel function of* $\{\phi, W(s) s \leqq t, \mu([0, s] \times B) \; s \leqq t \; B \in \sigma(\Gamma)\}$ *and there exists a constant* N, *depending only on* T *and* K *in* (2.3), *such that*

$$E\left\{\sup_{t \leqq T} \|q(t)\|_{l,0}^2\right\} \leqq N\left\{\|\phi\|_{l,0}^2 + \sup_{y,u} \sum_{i=0}^{d} \|\partial_i f^i(\cdot, y, u)\|_{l,0}^2 + \sup_y \|g(\cdot, y)\|_{l+1,0}^2\right\}, \tag{2.13}$$

$$l = 0, 1, \cdots, m.$$

(II) *Besides* (A.1)$\sim$(A.3), *we assume* $(A.4)_{l,r}$ *and* $\phi \in H_r^l$. *Then the following evaluation holds.*

$$E\left\{\sup_{t \leqq T} \|q(t)\|_{l,r}^2\right\} \leqq N'\left\{\|\phi\|_{l,r}^2 + \sup_{y,u} \sum_{i=0}^{d} \|\partial_i f^i(\cdot, y, u)\|_{l,r}^2 + \sup_y \|g(\cdot, y)\|_{l+1,r}^2\right\} \tag{2.14}$$

where $N' = N'(T, K, r)$.

(III) *Suppose*

$$F^i : [0, T] \times \mathbb{R}^d \times \Omega \to \mathbb{R}^1, \qquad i = 0, 1, \cdots, d,$$

and

$$G_k : [0, T] \times \mathbb{R}^d \times \Omega \to \mathbb{R}^1, \qquad k = 1, \cdots, d,$$

are $\mathcal{F}_t$*-adapted and*

$$E\left\{\int_0^T \|F^i(t)\|_{m+1,0}^2 \, dt\right\} < \infty, \qquad E\left\{\int_0^T \|G_k(t)\|_{m+1,0}^2 \, dt\right\} < \infty.$$

Let ξ be a solution of the following SPDE:

$$\text{(2.15)}\qquad \begin{cases} d\xi(t) = \sum_{i,j=0}^{d} \partial_i(\tilde{a}^{ij}(t, y+W(t), \mu)\partial_j\xi(t) + F^i(t))\,dt \\ \qquad\qquad + \left(\sum_{i=0}^{d} b^i(t, y+W(t))\partial_i\xi(t) + G(t)\right) dW(t) \\ \xi(0) = \varphi \in H^m. \end{cases}$$

Then, ξ satisfies the following evaluation.

$$\text{(2.16)}\qquad E\left\{\sup_{t\leqq T} \|\xi(t)\|_{l,0}^2\right\} \leqq N\Bigg(\|\varphi\|_{l,0}^2 + E\left\{\int_0^T \left(\sum_{i=0}^{d} \|\partial_i F^i(t)\|_{l,0}^2 + \sum_{k=1}^{d'} \|G_k(t)\|_{l+1,0}^2\right) dt\right\}\Bigg)$$

$(l = 0, 1, \cdots, m)$ *where* $N = N(T, K)$.

Remark. Krylov and Rozovskii proved Theorem 2.1, replacing (A.2) by a weaker condition

$$\text{(A.2')}\qquad a^{ij} = a^{ji}, \qquad i, j = 1, \cdots, d$$

and $(a^{ij} - \frac{1}{2}b^i \cdot b^j)_{i,j=1,\cdots,d}$ is a nonnegative definite matrix.

But we state all of our theorems under the condition (A.2), since we need (A.2) for Proposition 3.1, etc.

3. Continuous dependence of $q(\ , \phi, y, \mathscr{R})$ on ϕ, y, $\mathscr{R}$. Since we are mainly concerned with the probability law $\pi(\mathscr{R})$, we may assume the following canonical form, if necessary:

$\Omega = C(0, T; \mathbb{R}^{d'}) \times \Lambda$, $\mathscr{F} = \sigma(\Omega)$ = the topological σ-field on Ω;

W = the first coordinate function on Ω, $W(t, \omega) = W(\omega)(t)$;

μ = the second coordinate function on Ω,

$\mu(B, \omega) = \mu(\omega)(B)$, $B \in \sigma([0, T] \times \Gamma)$;

$\mathscr{F}_t = \sigma\{W(s), s \leqq t, \mu(B_1 \times B_2), B_1 \in \mathbf{B}[0, t], B_2 \in \sigma(\Gamma)\}$;

$P = \pi(\mathscr{R})$.

First we see the following lemma, which is crucial to the SPDE with ellipticity condition (A.2). It will be proved in the Appendix, according to [10].

LEMMA (*special case of Lemma* 2.1 *of* [10]). *For any $t \in [0, T]$, $y \in \mathbb{R}^{d'}$ and $\lambda \in \Lambda$, put $a^{ij}(\cdot) = a^{ij}(t, \cdot, y, \lambda)$ and $b^i(\cdot) = b^i(\cdot, y)$, for simplicity. Under the conditions* (A.1) *and* (A.2), *there exists a constant N, depending only on K in* (2.3), *T, and l* $(= 0, 1, \cdots, m)$, *such that*

$$(*)\qquad \int \sum_{|\gamma|\leqq l} \left(2\sum_{i,j=0}^{d} (-1)^{|i|} D^\gamma \partial_i u D^\gamma (a^{ij}\partial_j u + \hat{f}^i) + 3\left|D^\gamma\left(\sum_{i=0}^{d} b^i \partial_i u + \hat{g}\right)\right|^2\right) dx \leqq N\left\{\|u\|_l^2 + \sum_{i=0}^{d} \|\partial_i \hat{f}^i\|_l^2 + \sum_{k=0}^{d'} \|\hat{g}_k\|_{l+1}^2\right\}$$

for any fixed three functions u, $\hat{f}^i$, $\hat{g}_k \in H^{l+1}$ and $\hat{g} = (\hat{g}_1, \cdots, \hat{g}_{d'})$.

Remark. (1) When we take $\tilde{f}^i(\cdot, y, \mu)$ and $g(\cdot, y)$ of (2.1) as $\hat{f}^i$ and $\hat{g}$, respectively, (*) turns out to be in the following form:

$$2\langle \tilde{L}(t, y, \mu)u, u\rangle_l + 3|M(y)u|_l^2 \leqq N\left\{\|u\|_l^2 + \sum_{i=0}^{d} \|\partial_i \tilde{f}^i(\cdot, y, \mu)\|_l^2 + \sum_{k=0}^{d'} \|g_k(\cdot, y)\|_{l+1}^2\right\}.$$

(2) Reference [10] says that Lemma holds under conditions (A.1) and (A.2'), if we replace "3" of the integrand of the left-hand side with "1." So a stronger condition (A.2) yields a stronger evaluation (*), which is necessary for Proposition 3.1.

PROPOSITION 3.1. *There is a constant $C = C(T, K, l)$ such that*

$$\sup_{t \leqq T} E\{\|q(t)\|_l^4\} \leqq C\left\{\|\phi\|_l^4 + \sup_{y,u} \sum_{i=0}^{d} \|\partial_i f^i(\cdot, y, u)\|_l^4 + \sup_y \|g(\cdot, y)\|_{l+1}^4\right\}, \tag{3.1}$$

$$l = 0, 1, \cdots, m-1.$$

Proof. For simplicity, we put $\tilde{f}(t) = \tilde{f}(t, y + W(t), \mu)$, $g(t) = g(y + W(t))$, $\tilde{L}(t) = \tilde{L}(t, y + W(t), \mu)$, $M(t) = M(y + W(t))$ and $\langle , \rangle_l =$ duality pairing between H^{l-1} and H^{l+1} under $H^l = (H^l)^*$. Then q satisfies

$$(q(t), \eta)_l = (\phi, \eta)_l + \int_0^t \langle \tilde{L}(s)q(s), \eta\rangle_l \, ds + \int_0^t (M(s)q(s), \eta)_l \, dW(s) \tag{3.2}$$

$$\text{for } \eta \in H^{l+1}, \quad t \leqq T.$$

So Ito's formula derives

$$\begin{aligned} \|q(t)\|_l^2 &= \|\phi\|_l^2 + 2\int_0^t \langle \tilde{L}(s)q(s), q(s)\rangle_l \, ds \\ &\quad + \int_0^t \|M(s)q(s)\|_l^2 \, ds + 2\int_0^t (M(s)q(s), q(s))_l \, dW(s). \end{aligned} \tag{3.3}$$

Thus we see

$$\begin{aligned} &E[\|q(t)\|_l^4] - \|\phi\|_l^4 \\ &= 2E\left\{\int_0^t \|q(s)\|_l^2 \{2\langle \tilde{L}(s)q(s), q(s)\rangle_l + \|M(s)q(s)\|_l^2\} \, ds\right\} \\ &\quad + 4E\left\{\int_0^t (M(s)q(s), q(s))_l^2 \, ds\right\} \\ &\leqq 2E\left\{\int_0^t \|q(s)\|_l^2 \{2\langle \tilde{L}(s)q(s), q(s)\rangle_l + 3\|M(s)q(s)\|_l^2\} \, ds\right\} \\ &\leqq C_1 E\left[\int_0^t \|q(s)\|_l^2 \left\{\|q(s)\|_l^2 + \sum_{i=0}^{d} \|\partial_i \tilde{f}^i(s)\|_l^2 + \|g(s)\|_{l+1}^2\right\} ds\right] \end{aligned} \tag{3.4}$$

appealing to the lemma. Hence we have

$$\begin{aligned} E[\|q(t)\|_l^4] &\leqq C_2\left(E\left[\int_0^t \|q(s)\|_l^4 \, ds\right]\right. \\ &\quad \left. + \|\phi\|_l^4 + E\left[\int_0^t \left(\sum_{i=1}^{d} \|\partial_i \tilde{f}^i(s)\|_l^4 + \|g(s)\|_{l+1}^4\right) ds\right]\right). \end{aligned} \tag{3.5}$$

So Gronwall's inequality completes the proof. □

Now we will study continuous dependence of $q(\cdot, \phi, y, \mathcal{R})$ on $\mathfrak{R}$. For the following Theorem 3.1, we endow with the weak topology on $L^2(0, T; H^m)$ and H^{m-1}. Later Theorem 3.2 is concerned with strong topology on these spaces.

From now on, we always assume $m \geqq 3$.

THEOREM 3.1. *Suppose $\mathcal{R}_n \to \mathcal{R}$. Then, for $\phi \in H^m$ and $y \in \mathbb{R}^d$, we have*

$$(W_n, \mu_n, q(\cdot, \phi, y, \mathcal{R}_n)) \to (W, \mu, q(\cdot, \phi, y, \mathcal{R})) \text{ in law as } C(0, T; \mathbb{R}^{d'}) \times \Lambda \times [w - L^2(0, T; H^m)] - \text{r.v.} \tag{3.6}$$

$$(W_n, \mu_n, q(t, \phi, y, \mathcal{R}_n)) \to (W, \mu, q(t, \phi, y, \mathcal{R})) \text{ in law as } C(0, T; \mathbb{R}^{d'}) \times \Lambda \times [w - H^{m-1}] - \text{r.v.}, \tag{3.7}$$

where $w - X$ denotes the space X carrying the weak topology.

Proof. This theorem is an extension of Theorem 3.1 in [14] to the elliptic case (A.2) and we can apply the same method as [14], using the evaluation (*). First we introduce two spaces $\mathcal{H}_\gamma(D)$ and $\mathcal{H}_\gamma(D, T)$. Let D be a bounded open set of $\mathbb{R}^d$, with smooth boundary. Define $\mathcal{H}_\gamma(D)$ and $\mathcal{H}_\gamma(T, D)$ as follows (cf. [13]).

$$\mathcal{H}_\gamma(D) = \left\{ \varphi \in L^2(-\infty, \infty; H^{m-1}(D)); \int_{-\infty}^{\infty} |\tau|^{2\gamma} \|\hat{\varphi}(\tau)\|_*^2 \, d\tau < \infty \right\} \tag{3.8}$$

with the norm

$$\|\varphi\|_{\mathcal{H}_\gamma(D)}^2 = \int_{-\infty}^{\infty} \|\varphi(t)\|_{H^{m-1}(D)}^2 \, dt + \int_{-\infty}^{\infty} |\tau|^{2\gamma} \|\hat{\varphi}(\tau)\|_*^2 \, d\tau \tag{3.9}$$

where, for simplicity, we put $\hat{\varphi}(\tau) = \int_{-\infty}^{\infty} \exp(-2\pi i \tau t) \varphi(t) \, dt$ in this proof and $\|\cdot\|_* =$ norm of $(H^{m-1}(D))^*$ (= dual space of $H^{m-1}(D)$ under $H^{m-2}(D) = (H^{m-2}(D))^*$) and

$$\mathcal{H}_\gamma(T, D) = \{\varphi|_{[0,T]}; \varphi \in \mathcal{H}_\gamma(D)\} \tag{3.10}$$

with the norm

$$\|\varphi\|_{\mathcal{H}_\gamma(T,D)} = \inf\{\|\psi\|_{\mathcal{H}_\gamma(D)}; \varphi = \psi \text{ a.e. on } [0, T]\}, \tag{3.11}$$

respectively. □

Now we divide the proof into three steps. The first step is the preliminary lemma, which is useful for proving the compactness of space of solutions.

LEMMA 3.1. *For any fixed $\gamma \in (0, \frac{1}{4})$,*

$$q(\ , \phi, y, \mathcal{R}) \in \mathcal{H}_\gamma(T, D), \quad \text{w.p. 1} \tag{3.12}$$

holds, and there is a constant $K_1 = K_1(T, K)$, such that

$$E\{\|q(\ , \phi, y, \mathcal{R})\|_{\mathcal{H}_\gamma(T,D)}^2\} \leqq K_1 I_m(\phi, f, g), \quad \text{for } \forall \mathcal{R} \in \mathfrak{R}, \tag{3.13}$$

where

$$I_m(\phi, f, g) = \|\phi\|_m^2 + \sup_{y,u} \sum_{i=0}^{d} \|\partial_i f^i(\cdot, y, u)\|_m^2 + \sup_y \|g(\cdot, y)\|_{m+1}^2. \tag{3.14}$$

Proof. Put

$$h(t) = \begin{cases} h(t, \mathcal{R}), & t \in [0, T] \\ 0, & t \notin [0, T], \end{cases}$$

for $h(t, \mathcal{R}) = q(t, \phi, y, \mathcal{R})$, $f(\cdot, y + W(t), \mu)$ and $g(\cdot, y + W(t))$. $\tilde{L}(t)$ and $M(t)$, $t \in (-\infty, \infty)$, are defined in the same way as (2.5) and (2.2), respectively. Since q is a solution, the following equality (3.15) holds, for any $\eta \in H^m$,

$$\begin{aligned} (q(t), \eta)_{m-1} &= (\phi, \eta)_{m-1} + \int_0^t \langle \tilde{L}(s) q(s), \eta \rangle_{m-1} \, ds \\ &\quad + \int_0^t (M(s) q(s), \eta)_{m-1} \, dW(s). \end{aligned} \tag{3.15}$$

Therefore, we have

$$2\pi i\tau(\hat{q}(\tau),\eta)_{m-1}=\int_{-\infty}^{\infty}\left(-\frac{d}{dt}\exp(-2\pi i\tau t)\right)(q(t),\eta)_{m-1}\,dt$$

$$=(\phi,\eta)_{m-1}-\exp(-2\pi i\tau T)(q(T),\eta)_{m-1}+\langle\widehat{\tilde{L}q}(\tau),\eta\rangle_{m-1}$$

$$+\int_0^T\exp(-2\pi i\tau t)(M(t)q(t),\eta)_{m-1}\,dW(t). \tag{3.16}$$

Let $\eta_j\in C_0^\infty(\mathbb{R}^d)$, $j=1,2,\cdots$ be a complete orthonormal system of H^m. Then we get

$$4\pi^2\tau^2E[\|\hat{q}(\tau)\|_{m-2}^2]=4\pi^2\tau^2\sum_{j=1}^{\infty}E\{|(\hat{q}(\tau),\eta_j)_{m-1}|^2\}$$

$$\leqq C_1\Big(\|\phi\|_{m-2}^2+E\{\|q(T)\|_{m-2}^2\}+E\{\|\tilde{L}q(\tau)\|_{m-2}^2\}$$

$$+\sum_{k=1}^{d'}\int_0^T\|M_k(t)q(t)\|_{m-2}^2\,dt\Big). \tag{3.17}$$

Since $\|\cdot\|_*\leqq\|\cdot\|_{m-2}\leqq\|\cdot\|_{m-1}$, we see

$$\tau^2E[\|\hat{q}(\tau)\|_*^2]\leqq C_2\Big(\|\phi\|_{m-1}^2+E[\|q(T)\|_{m-1}^2]$$

$$+E\left[\int_0^T(\|q(t)\|_m^2+\|g(t)\|_{m-1}^2)\,dt\right]+E[\|\widehat{\tilde{L}q}(\tau)\|_{m-2}^2]\Big)$$

$$\leqq C_3\left\{I_m(\phi,f,g)+E[\|\widehat{\tilde{L}q}(\tau)\|_{m-2}^2]\right\}.$$

Hence, for any fixed $\kappa\in(1,\frac{3}{2})$,

$$\int_{-\infty}^{\infty}E[|\tau|^{2\gamma}\|\hat{q}(\tau)\|_*^2]\,d\tau$$

$$\leqq\int_{|\tau|\leqq1}E[\|\hat{q}(\tau)\|_*^2]\,d\tau+\int_{|\tau|>1}E\left[\frac{2|\tau|^2}{1+|\tau|^\kappa}\|\hat{q}(\tau)\|_*^2\right]d\tau$$

$$\leqq C_4\Big(\int_{-\infty}^{\infty}E[\|\hat{q}(\tau)\|_{m-2}^2]\,d\tau+I_m(\phi,y,\mathcal{R})\int_{-\infty}^{\infty}\frac{2}{1+|\tau|^\kappa}\,d\tau$$

$$+\int_{-\infty}^{\infty}E[\|\widehat{\tilde{L}q}(\tau)\|_{m-2}^2]\,d\tau\Big)$$

$$\leqq C_5\Big(\int_{-\infty}^{\infty}E\{\|q(t)\|_m^2+\|\tilde{L}q(t)\|_{m-2}^2\}\,dt+I_m(\phi,f,g)\Big)$$

$$\leqq C_6I_m(\phi,f,g)$$

where $C_i=C_i(T,K)$. From this we get

$$E[\|q\|_{\mathcal{H}_\gamma(D)}^2]\leqq K_1I_m(\phi,f,g), \tag{3.18}$$

and complete the proof of Lemma 3.1. □

Second step. Let D_k $(k=1, 2, \cdots)$ be a bounded and open subset of $\mathbb{R}^d$ with smooth boundary, $\bar{D}_k \subset D_{k+1}$ and $\cup_{k=1}^{\infty} D_k = \mathbb{R}^d$. Define a metric d by

$$d(p, q) = \sum_{k=1}^{\infty} \frac{1}{2^k} \min\left(1, \left\{\int_0^T \| p(t) - q(t) \|_{H^{m-2}(D_k)^2} \, dt\right\}^{1/2}\right)$$

for $p, q \in L^2(0, T; H^{m-2})$. $\mathcal{H}^{m-2}(0, T)$ denotes the completion of $L^2(0, T; H^{m-2})$ with respect to the metric d. Put $S_1 = C(0, T; \mathbb{R}^{d'}) \times \Lambda \times \mathcal{H}^{m-2}(0, T)$ and $S_2 = C(0, T; \mathbb{R}^{d'}) \times \Lambda \times [w - L^2(0, T; H^{m-2})]$. For $\mathcal{R} = (W, \mu)$, $m_1(\mathcal{R})$ and $m_2(\mathcal{R})$ denote the image measures of $(W, \mu, q(\cdot, \phi, y, \mathcal{R}))$ on S_1 and S_2, respectively. $B_r = \{q \in L^2(0, T; H^{m-2});\ \|q\|_{\mathcal{H}_\gamma(T, D_k)} \leqq (2^k r)^{1/2},\ k = 1, 2, \cdots\}$ is compact in $\mathcal{H}^{m-2}(0, T)$, because the injection $\mathcal{H}_\gamma(T, D_k) \to L^2(0, T; H^{m-2}(D_k))$ is a compact operator (cf. [13]).

On the other hand, Lemma 3.1 asserts

$$P(q(\cdot, \phi, y, \mathcal{R}) \notin B_r) \leqq K_1 I_m(\phi, f, g)/r.$$

Hence, $\{m_1(\mathcal{R}), \mathcal{R} \in \mathfrak{R}\}$ is relatively compact by Proposition 2.3. Moreover, $\{m_2(\mathcal{R}), \mathcal{R} \in \mathfrak{R}\}$ is also relatively compact by (2.13) and Remark 3.3 in [14].

Third step. Suppose $\mathcal{R}_n \to \mathcal{R}$. Then we can choose a subsequence $\{n_j\}$, such that $m_1(\mathcal{R}_{n_j})$ and $m_2(\mathcal{R}_{n_j})$ converge to some probability measures m_1 and m_2, respectively. So their marginal distributions on $C(0, T; \mathbb{R}^{d'}) \times \Lambda$ coincide with $\pi(\mathcal{R})$ and $i_k(m_1(\mathcal{R}_n)) = j_k(m_2(\mathcal{R}_n))$ and $i_k(m_1) = j_k(m_2)$ $(k = 1, 2, \cdots)$, where

$$i_k : S_1 \to C(0, T; \mathbb{R}^{d'}) \times \Lambda \times L^2(0, T; H^{m-2}(D_k))$$

and

$$j_k : S_2 \to C(0, T; \mathbb{R}^{d'}) \times \Lambda \times L^2(0, T; H^{m-2}(D_k))$$

are the canonical injections.

$$m_1(C(0, T; \mathbb{R}^{d'}) \times \Lambda \times L^2(0, T; H^{m-2})) = 1$$

holds by (2.13).

Endowing with the metric d, we can apply Skorokhod's theorem. Hence, there exist S_1-valued random variables $(\hat{W}_{n_j}, \hat{\mu}_{n_j}, \hat{q}_{n_j})$ and $(\hat{W}, \hat{\mu}, \hat{q})$ on a suitable probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$, such that

(3.19) the law of $(\hat{W}_{n_j}, \hat{\mu}_{n_j}, \hat{q}_{n_j}) = m_1(\mathcal{R}_{n_j})$

The law of $(\hat{W}, \hat{\mu}, \hat{q}) = m_1$ (= limit measure of $m_1(\mathcal{R}_{n_j})$),

(3.20) with probability 1,

(I) $\hat{W}_{n_j} \to \hat{W}$ uniformly on $[0, T]$

(II) $\hat{\mu}_{n_j} \to \hat{\mu}$ weakly

(III) $\hat{q}_{n_j} \to \hat{q}$ in $\mathcal{H}^{m-2}(0, T)$.

Moreover, since (3.1) implies the uniform integrability, we have

(IV) $\hat{q}_{n_j}|_{D_k} \to \hat{q}|_{D_k}$ in $L^2([0, T] \times \Omega; H^{m-2}(D_k))$ for $k = 1, 2, \cdots$.

Hence, from (I) and (II), we see, for all $x \in \mathbb{R}^d$

$$\begin{aligned}
&\int_0^T \psi(t) \tilde{a}(t, x, y + \hat{W}_{n_j}(t), \hat{\mu}_{n_j}) \, dt \\
&\quad = \int_0^T \psi(t) \int_\Gamma a(x, y + \hat{W}_{n_j}(t), u) \hat{\mu}'_{n_j}(t, du) \, dt \\
&\quad \xrightarrow[n_j \to \infty]{} \int_0^T \psi(t) \int_\Gamma a(x, y + \hat{W}(t), u) \hat{\mu}'(t, du) \, dt \\
&\quad = \int_0^T \psi(t) \tilde{a}(t, x, y + \hat{W}(t), \hat{\mu}) \, dt
\end{aligned} \tag{3.21}$$

for any bounded continuous function ψ on $[0, T]$. Namely, we have

$$\tilde{a}(\cdot, x, y+\hat{W}_{n_j}, \hat{\mu}_{n_j}) \to \tilde{a}(\cdot, x, y+\hat{W}, \hat{\mu}) \quad \text{in } [w-L^2(0, T)]. \tag{3.22}$$

Since $\tilde{q}_{n_j}$ is a response for $\mathcal{R}_{n_j}=(\hat{W}_{n_j}, \hat{\mu}_{n_j})$, we see, for any bounded absolutely continuous function ψ with $\psi' \in L^2(0, T)$ and $\psi(T)=0$, and $\eta \in C_0^\infty$,

$$\begin{aligned}
&\int_0^T \psi(t) d(\hat{q}_{n_j}(t), \eta) \\
&\quad = -\psi(0)(\phi, \eta) - \int_0^T (\hat{q}_{n_j}(t), \eta)\psi'(t)\, dt \\
&\quad = \int_0^T \langle \tilde{L}(t, y+\hat{W}_{n_j}(t), \hat{\mu}_{n_j})\hat{q}_{n_j}(t), \eta\rangle \psi(t)\, dt \\
&\qquad + \int_0^T \psi(t)(M(y+\hat{W}_{n_j}(t))\hat{q}_{n_j}(t), \eta)\, d\hat{W}_{n_j}(t).
\end{aligned} \tag{3.23}$$

Hence, we get, as $n_j \to \infty$

$$\begin{aligned}
&-\int_0^T (\hat{q}(t), \eta)\psi'(t)\, dt \\
&\quad = \psi(0)(\phi, \eta) + \int_0^T \langle \tilde{L}(t, y+\hat{W}(t), \hat{\mu})\hat{q}(t), \eta\rangle \psi(t)\, dt \\
&\qquad + \int_0^T \psi(t)(M(y+\hat{W}(t))\hat{q}(t), \eta)\, d\hat{W}(t)
\end{aligned} \tag{3.24}$$

whenever supp $\eta \subset D_k$ for some k.

Equation (3.24) yields that $\hat{q}$ is a response for $(\hat{W}, \hat{\mu})$. Since $\pi(\hat{W}, \hat{\mu}) = \pi(\mathcal{R})$, we obtain

$m_1 =$ the law of $(\hat{W}, \hat{\mu}, \hat{q}) = m_1(\mathcal{R})$ and also $m_2 = m_2(\mathcal{R})$. This fact concludes (3.6).

In the same way we can prove (3.7). □

Now we will deal with $L^2(0, T; H^{m-2})$ and H^{m-2} instead of $[w-L^2(0, T; H^{m-2})]$ and $[w-H^{m-2}]$. Put $\Phi_r = H^m \cap H_r^{m-2}$, $r>0$, with the norm $|\cdot|_r = \|\cdot\|_m + \|\cdot\|_{m-2,r}$. By applying [11], we evaluate $q(t, x)$ for large $|x|$.

THEOREM 3.2. *Suppose* $(A.4)_{m-2,r}$ *besides* (A.1)$\sim$(A.3). *Then for* $\phi \in \Phi_r$, *we have*

$$q(\cdot, \phi, y, \mathcal{R}_n) \to q(\cdot, \phi, y, \mathcal{R}) \quad \textit{in law as } L^2(0, T; H^{m-2}) - \text{r.v.}, \tag{3.25}$$

and for any fixed t

$$q(t, \phi, y, \mathcal{R}_n) \to q(t, \phi, y, \mathcal{R}) \quad \textit{in law as } H^{m-2} - \text{r.v.}, \tag{3.26}$$

whenever $\mathcal{R}_n \to \mathcal{R}$.

Proof. By Theorem 2.1, there exists a constant C depending only on T, K, r, and ϕ such that

$$E\left[\int_{\mathbb{R}^d} (1+|x|^2)^r \{D^\alpha q(t, x, \phi, y, \mathcal{R})\}^2\, dx\right] \leqq C \tag{3.27}$$

for all t, α; $0 \leqq t \leqq T$, $0 \leqq |\alpha| \leqq m-2$, and $\mathcal{R} \in \mathfrak{R}$. Hence, we have

$$E\left[\int_{|x|>\rho} D^\alpha q(t, x)^2\, dx\right] \leqq \frac{C}{(1+\rho^2)^r}. \tag{3.28}$$

By virtue of Skorokhod's theorem, there exist $L^2(0, T; H^{m-2})$-valued random variables $\hat{q}_n$ and $\hat{q}$ on a suitable probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$, such that

(3.29) $\hat{q}_n$ and $\hat{q}$ have the same laws as $q(\ , \phi, y, \mathcal{R}_n)$ and $q(\ , \phi, y, \mathcal{R})$, respectively,

and with probability 1

$$\hat{q}_n \to \hat{q} \text{ in } L^2(0, T; H^{m-2}(D)) \tag{3.30}$$

for any bounded subset D of $\mathbb{R}^d$.

On the other hand, we see from (3.1)

$$E\left[\left(\int_D \{D^\alpha q(t, x)\}^2 \, dx\right)^2\right] \leqq E[\|q(t)\|_{m-2}^4] \leqq C' \qquad \text{for } 0 \leqq |\alpha| \leqq m-2, \tag{3.31}$$

where C' is independent from D, t, and $\mathcal{R}$.

Since this implies the uniform integrability, we get

$$E\left[\int_0^T \int_D \sum_{|\alpha| \leqq m-2} \{D^\alpha \hat{q}_n(t, x) - D^\alpha \hat{q}(t, x)\}^2 \, dx \, dt\right] \to 0. \tag{3.32}$$

Combining (3.32) with (3.28), we obtain

$$E\left[\int_0^T \|\hat{q}_n(t) - \hat{q}(t)\|_{m-2}^2 \, dt\right] \to 0.$$

This concludes (3.25).

For the proof of (3.26), we can apply the same argument. □

Putting

$$\Phi = \bigcup_{r>0} \Phi_r, \tag{3.33}$$

we see the following corollary.

COROLLARY 3.1. *Suppose* $\mathcal{R}_n = (W_n, \mu_n)$ *tends to* $\mathcal{R} = (W, \mu)$. *Then, under the conditions* (A.1) ~ (A.3), $(A.4)_{m-2}$ *and* $\phi \in \Phi$, *there exist* $\hat{\mathcal{R}}_n = (\hat{W}_n, \hat{\mu}_n)$ *and* $\hat{\mathcal{R}} = (\hat{W}, \hat{\mu})$, *on a suitable probability space, such that*

(I) $\pi(W_n, \mu_n) = \pi(\hat{W}, \hat{\mu}_n)$, $\pi(W, \mu) = \pi(\hat{W}, \hat{\mu})$

and with probability 1,

(II) $\hat{W}_n \to \hat{W}$ *uniformly on* $[0, T]$

(III) $\hat{\mu}_n \to \hat{\mu}$ *weakly*

(IV) $\hat{q}_n \to \hat{q}$ *in* $L^2(0, T; H^{m-2})$

(V) $\hat{q}_n(t) \to \hat{q}(t)$ *in* H^{m-2}

where $\hat{q}_n$ *and* $\hat{q}$ *are responses for* $\hat{\mathcal{R}}_n$ *and* $\hat{\mathcal{R}}$, *respectively.*

Next we will study the dependence of q on the initial (ϕ, y).

THEOREM 3.3.

$$E[\sup_{t \leqq T} \|q(t, \phi, y, \mathcal{R}) - q(t, \psi, y, \mathcal{R})\|_l^2] \leqq N \|\phi - \psi\|_l^2 \tag{3.34}$$

$(l = 0, 1, \cdots, m)$, *where* N *is the constant of* (2.13).

$$E[\sup_{t \leqq T} \|q(t, \phi, y_1, \mathcal{R}) - q(t, \phi, y_2, \mathcal{R})\|_l^2] \leqq N_1(1 + I_{l+2}(\phi, f, g))|y_1 - y_2|^2) \tag{3.35}$$

$(l = 0, 1, \cdots, m-2)$, *where* $N_1 = N_1(T, K)$.

Proof. Put $p=q(\cdot, \phi, y, \mathcal{R})-q(\cdot, \psi, y, \mathcal{R})$. Then p satisfies the following SPDE

$$\begin{aligned} dp(t) &= \sum_{i,j=0}^{d} \partial_i(\tilde{a}^{ij}(t, y+W(t), \mu)\partial_j p(t)\, dt \\ &\quad + \sum_{i=0}^{d} b^i(t, y+W(t))\partial_i p(t)\, dW(t) \\ p(0) &= \phi-\psi. \end{aligned} \tag{3.36}$$

Therefore, (2.13) derives (3.34).

Put $\xi=q_1-q_2$ where $q_i=q(\cdot, \phi, y_i, \mathcal{R})$. Then we have

$$\begin{aligned} d\xi(t) &= \{\tilde{L}_1(t)\xi(t)+(\tilde{L}_1(t)-\tilde{L}_2(t))q_2(t)\}\, dt \\ &\quad +\{M_1(t)\xi(t)+(M_1(t)-M_2(t))q_2(t)\}\, dW(t) \\ \xi(0) &= 0 \end{aligned} \tag{3.37}$$

where $\tilde{L}_i(t)=\tilde{L}(t, y_i+W(t), \mu)$ and $M_i(t)=M(t, y_i+W(t))$, $i=1, 2$. So (2.16) asserts

$$\begin{aligned} E\left[\sup_{t\leqq T}\|\xi(t)\|_l^2\right] \leqq NE\Bigg[&\int_0^T \|(\tilde{L}_i(t)-\tilde{L}_2(t))q_2(t)\|_l^2\, dt \\ &+\int_0^T \|(M_1(t)-M_2(t))q_2(t)\|_{l+1}^2\, dt\Bigg]. \end{aligned} \tag{3.38}$$

Thus we see, from (A.1),

$$E\left[\sup_{t\leqq T}\|\xi(t)\|_l^2\right] \leqq N_1|y_1-y_2|^2(1+I_{l+2}(\phi, f, g)) \tag{3.39}$$

$(l=0, 1, \cdots, m-2)$, where $N_1=N_1(T, K)$. □

COROLLARY 3.2. *There is a constant $N_2=N_2(T, K)$ such that*

$$\begin{aligned} &E\left[\sup_{t\leqq T}\|q(t, \phi_1, y_1, \mathcal{R})-q(t, \phi_2, y_2, \mathcal{R})\|_l^2\right] \\ &\leqq N_2\Bigg(|y_1-y_2|^2\Bigg\{1+\min(\|\phi_1\|_{l+2}^2, \|\phi_2\|_{l+2}^2)+\sup_y\|g(\cdot, y)\|_{l+3}^2 \\ &\quad +\sup_{y,u}\sum_{i=0}^{d}\|\partial_i f(\cdot, y, u)\|_{l+2}^2\Bigg\}+\|\phi_1-\phi_2\|_l^2\Bigg) \qquad l=0, 1, \cdots, m-2. \end{aligned} \tag{3.40}$$

4. Optimal relaxed systems. Let $F: L^2(0, T; H^{m-2})\to\mathbb{R}^1$ and $G: H^{m-2}\to\mathbb{R}^1$ be uniformly continuous with linear growth, namely

(4.1) for any $\varepsilon>0$, there is $\delta=\delta(\varepsilon)>0$ such that

$$\begin{aligned} &|F(\psi_1)-F(\psi_2)|<\varepsilon \quad \text{if } \|\psi_1-\psi_2\|_{L^2(0,T;H^{m-2})}<\delta \\ &|G(\varphi_1)-G(\varphi_2)|<\varepsilon \quad \text{if } \|\varphi_1-\varphi_2\|_{m-2}<\delta \end{aligned}$$

and there is $\alpha>0$ such that

$$\begin{aligned} |F(\psi)| &\leqq \alpha(1+\|\psi\|_{L^2(0,T;H^{m-2})}) \\ |G(\varphi)| &\leqq \alpha(1+\|\varphi\|_{m-2}). \end{aligned} \tag{4.2}$$

For $\mathcal{R}\in\mathfrak{R}$, we will define the pay-off function J and the value function V by

$$J(\phi, y, \mathcal{R})=E[F(q(\cdot, \phi, y, \mathcal{R}))+G(q(T, \phi, y, \mathcal{R}))] \tag{4.3}$$

and

$$V(\phi, y) = \inf_{\mathcal{R} \in \mathfrak{R}} J(\phi, y, \mathcal{R}),$$

respectively. Then, Theorem 3.2 and Proposition 2.3 assert the existence of an optimal relaxed system. Now we have Theorem 4.1.

THEOREM 4.1. *Under the conditions* (A.1)$\sim$(A.3) *and* $(A.4)_{m-2}$, *there exists an optimal relaxed system* $\mathcal{R}^* = \mathcal{R}^*(\phi, y)$ *for* $\phi \in \Phi$ (*see* (3.33)), *namely,*

$$V(\phi, y) = J(\phi, y, \mathcal{R}^*) \tag{4.4}$$

holds. Moreover, for any $r > 0$, *we can choose* $\mathcal{R}^*(\phi, y)$, *so that* $\pi(\mathcal{R}^*(\phi, y))$ *is a Borel map from* $\Phi_r \times \mathbb{R}^d$ *into* $\mathcal{P}(C[0, T] \times \Lambda)$.

Proof. Suppose $\mathcal{R}_n \to \mathcal{R}$. Putting $q_n = q(\cdot, \phi, y, \mathcal{R}_n)$ and $q = q(\cdot, \phi, y, \mathcal{R})$, $F(q_n)$ and $G(q_n(T))$ converge to $F(q)$ and $G(q(T))$ in law, respectively. On the other hand, (3.1) derives

$$\sup_n E[F(q_n)^2] \leqq C_1\left(1 + \sup_n E\left[\int_0^T \|q_n(t)\|_{m-2}^2 \, dt\right]\right) < \infty.$$

Thus, the uniform integrability asserts

$$E[F(q_n)] \to E[F(q)].$$

In the same way, we can prove

$$E[G(q_n(T))] \to E[G(q(T))].$$

Hence, $J(\phi, y, \mathcal{R})$ is continuous in $\mathcal{R}$. Thus, Proposition 2.3 concludes (4.4).

For the proof of the latter half, we apply the same arguments as in Chapter 12 of [19]. Putting

$$\mathfrak{X}(\phi, y) = \{\pi(\mathcal{R});\ V(\phi, y) = J(\phi, y, \mathcal{R})\}, \tag{4.5}$$

we show the following lemma.

LEMMA 4.1. $\mathfrak{X}(\phi, y)$ *is nonempty and compact.*

Proof. $\mathfrak{X}(\phi, y)$ is nonempty by (4.4). So we will prove the closedness of $\mathfrak{X}(\phi, y)$. Suppose $\pi(\mathcal{R}_n) \in \mathfrak{X}(\phi, y)$ and converges to $\pi(\mathcal{R})$ weakly. Then $J(\phi, y, \mathcal{R}_n) \to J(\phi, y, \mathcal{R})$.

Hence $J(\phi, y, \mathcal{R}) = V(\phi, y)$, namely, "$\pi(\mathcal{R}) \in \mathfrak{X}(\phi, y)$." □

Let $\phi_n \to \phi$ in Φ_r and $y_n \to y$. Suppose $\pi(\mathcal{R}_n) \in \mathfrak{X}(\phi_n, y_n)$ and $\pi(\mathcal{R}_n) \to \pi(\mathcal{R})$ weakly. Then we will show $\pi(\mathcal{R}) \in \mathfrak{X}(\phi, y)$, which completes the proof.

$$\begin{aligned} &|J(\phi_n, y_n, \mathcal{R}_n) - J(\phi, y, \mathcal{R})| \\ &\quad \leqq |J(\phi_n, y_n, \mathcal{R}_n) - J(\phi, y, \mathcal{R}_n)| + |J(\phi, y, \mathcal{R}_n) - J(\phi, y, \mathcal{R})|. \end{aligned} \tag{4.6}$$

We see, from (3.1), (3.40) and (4.1), the following:

$$\begin{aligned} &\text{1st term of the right-hand side of (4.6)} \\ &\quad \leqq 2\varepsilon + E[|F(q(\phi_n, y_n, \mathcal{R}_n)) - F(q(\phi, y, \mathcal{R}_n))|;\ A_n] \\ &\qquad + E[|G(q(T, \phi_n, y_n, \mathcal{R}_n)) - G(q(T, \phi, y, \mathcal{R}_n))|;\ B_n] \\ &\quad \leqq 2\varepsilon + C_1(1 + \|\phi_n\|_{m-2} + \|\phi\|_{m-2})\{|y_n - y|(1 + \|\phi\|_m) + \|\phi_n - \phi\|_{m-2}\}/\delta \end{aligned} \tag{4.7}$$

with C_1 independent from ε and n, where

$$A_n = \{\|q(\phi_n, y_n, \mathcal{R}_n) - q(\phi, y, \mathcal{R}_n)\|_{L^2(0,T;H^{m-2})} > \delta\}$$

and

$$B_n = \{\|q(T, \phi_n, y_n, \mathcal{R}_n) - q(T, \phi, y, \mathcal{R}_n)\|_{m-2} > \delta\}.$$

Since $J(\phi, y, \mathcal{R})$ is continuous in $\mathcal{R}$, (4.6) and (4.7) yield

$$J(\phi_n, y_n, \mathcal{R}_n) \to J(\phi, y, \mathcal{R}). \tag{4.8}$$

Using $|V(\phi_n, y_n) - V(\phi, y)| \leqq \sup_{\mathcal{R} \in \mathfrak{R}} |J(\phi_n, y_n, \mathcal{R}) - J(\phi, y, \mathcal{R})|$, (4.7) derives

$$V(\phi_n, y_n) \to V(\phi, y). \tag{4.9}$$

Thus, we have

$$J(\phi, y, \mathcal{R}) = \lim_{n\to\infty} J(\phi_n, y_n, \mathcal{R}_n) = \lim_{n\to\infty} V(\phi_n, y_n) = V(\phi, y)$$

Namely, $\pi(\mathcal{R}) \in \mathfrak{X}(\phi, y)$.

Therefore, we can take a Borel selector S_r of $\mathfrak{X}(\phi, y)$, i.e., $S_r : \Phi_r \times \mathbb{R}^{d'} \to \mathcal{P}(C[0, T] \times \Lambda)$, Borel map, such that $S_r(\phi, y) \in \mathfrak{X}(\phi, y)$ ([19, Chap. 12]).

So $S_r(\phi, y) = \pi(\mathcal{R}^*(\phi, y))$ holds. This completes the proof of Theorem 4.1. □

Since a relaxed control turns out to be an admissible control under the Roxin condition, we can get an optimal admissible control. Now we introduce the convexity condition for coefficients of (2.1). Put $c(y, u) = (a^{ij}(\cdot, y, u), f^i(\cdot, y, u); i, j = 0, \cdots, d)$ and $C(y, \Gamma) = \{c(y, u); u \in \Gamma\}$.

CONVEXITY CONDITION (*Roxin condition*). *For any $y \in \mathbb{R}^{d'}$, $C(y, \Gamma)$ is a convex subset of $C(\mathbb{R}^d; \mathbb{R}^{(d+1)(d+2)})$.*

Endowing with the compact uniform topology on $C(\mathbb{R}^d; \mathbb{R}^{(d+1)(d+2)})$ we have Proposition 4.2.

PROPOSITION 4.2. *Under the convexity condition, $C(y, \Gamma)$ is compact and convex.*

Proof. $c(y, \cdot)$ is continuous in Γ. Since Γ is compact, $C(y, \Gamma)$ is compact. □

Let us set $\tilde{c}(\cdot, y, \nu) = \int_\Gamma c(\cdot, y, u)\nu(du)$ for $\nu \in \mathcal{P}(\Gamma)$, namely, $\tilde{c}(\cdot, y, \nu) = (\tilde{a}(\cdot, y, \nu), \tilde{f}(\cdot, y, \nu))$. Putting $\Gamma(y, \nu) = \{u \in \Gamma; \tilde{c}(\cdot, y, \nu) = c(\cdot, y, u)\}$, we see Proposition 4.3.

PROPOSITION 4.3. *$\Gamma(y, \nu)$ is nonempty and compact.*

Proof. Since $C(y, \Gamma)$ is convex and compact, $\tilde{c}(\cdot, y, \nu) \in C(y, \Gamma)$. So $\Gamma(y, \nu) \neq \varnothing$. Now we will show that $\Gamma(y, \nu)$ is closed. Suppose $u_n \in \Gamma(y, \nu)$ and $u_n \to u$. Then $c(\cdot, y, u_n) \to c(\cdot, y, u)$. Thus $c(\cdot, y, u) = \tilde{c}(\cdot, y, \nu)$. This completes the proof. □

Again appealing to [19, Chap. 12], we see Proposition 4.4.

PROPOSITION 4.4. *There exists a Borel selector $\tilde{S}$ of $\Gamma(y, \nu)$, i.e., $\tilde{S} : \mathbb{R}^{d'} \times \mathcal{P}(\Gamma) \to \Gamma$ Borel map, such that $\tilde{S}(y, \nu) \in \Gamma(y, \nu)$.*

Proof. Suppose $\nu_n \to \nu$ weakly and $y_n \to y$. Then

$$\begin{aligned} &|\tilde{c}(x, y_n, \nu_n) - \tilde{c}(x, y, \nu)| \\ &\quad \leqq \int_\Gamma |c(x, y_n, u) - c(x, y, u)| d\nu_n + |\tilde{c}(x, y, \nu_n) - \tilde{c}(x, y, \nu)| \\ &\quad \leqq \sup_{x,u} |c(x, y_n, u) - c(x, y, u)| + |\tilde{c}(x, y, \nu_n) - \tilde{c}(x, y, \nu)| \end{aligned} \tag{4.10}$$

holds. By the uniform continuity of c, the first term tends to 0 as $n \to \infty$. The second term also tends to 0 by the assumption $\nu_n \to \nu$ weakly. Hence, as $n \to \infty$, $|\tilde{c}(\cdot, y, \nu_n) - \tilde{c}(\cdot, y, \nu)| \to 0$ uniformly in any compact set of $\mathbb{R}^d$, by virtue of uniform continuity of c. This derives

$$\tilde{c}(\cdot, y_n, \nu_n) \to \tilde{c}(\cdot, y, \nu), \quad \text{as} \quad n \to \infty. \tag{4.11}$$

Suppose $u_n \in \Gamma(y_n, \nu_n)$ tends to u. Since $c(\cdot, y_n, u_n) \to c(\cdot, y, u)$, (4.11) yields "$u \in \Gamma(y, \nu)$." This concludes the proof of Proposition 4.4. □

For $\mathcal{R} = (\Omega, \mathcal{F}, \mathcal{F}_t, P, W, \mu)$, we define an $\mathcal{F}_t$-adapted process U by

$$U(t) = \tilde{S}(y + W(t), \mu'(t)). \tag{4.12}$$

Then we have

$$\tilde{c}(x, y+W(t), \mu'(t)) = c(x, y+W(t), U(t)) \tag{4.13}$$

and

$$\tilde{L}(t, y+W(t), \mu) = L(y+W(t), U(t)). \tag{4.14}$$

Hence, $q = q(\ , \phi, y, \mathcal{R})$ satisfies

$$\begin{aligned} &dq(t) = L(y+W(t), U(t))q(t)\,dt + M(y+W(t))q(t)\,dW(t) \\ &q(0) = \phi. \end{aligned} \tag{4.15}$$

Since (4.15) has a unique solution, q turns out to be the response for the admissible system $\mathcal{A} = (\Omega, \mathcal{F}, \mathcal{F}_t, P, W, U)$.

Although an admissible system can be regarded as a relaxed system, we denote the pay-off function by $J(\phi, y, \mathcal{A})$, stressing an admissible system $\mathcal{A}$. Recalling Theorem 4.1, we get Theorem 4.2.

THEOREM 4.2. *Supposing* (A.1)$\sim$(A.3), $(A.4)_{m-1}$ *and the convexity condition, there is an optimal admissible system* $\mathcal{A}^*$, *for* $\phi \in \Phi$, *such that*

$$V(\phi, y) = \inf_{\mathcal{A} \in \mathcal{U}} J(\phi, y, \mathcal{A}) = J(\phi, y, \mathcal{A}^*). \tag{4.16}$$

Proof. Put $U^*(t) = \tilde{S}(y+W^*(t), \mu^{*\prime}(t))$ for an optimal relaxed system $\mathcal{R}^* = (\Omega, \mathcal{F}, \mathcal{F}_t, P^*, W^*, \mu^*)$. Then $\mathcal{A}^* = (\Omega, \mathcal{F}, \mathcal{F}_t, P^*, W^*, U^*)$ satisfies

$$V(\phi, y) = J(\phi, y, \mathcal{R}^*) = J(\phi, y, \mathcal{A}^*) \geqq \inf_{\mathcal{A} \in \mathcal{U}} J(\phi, y, \mathcal{A}) \tag{4.17}$$

Since $V(\phi, y) \leqq \inf_{\mathcal{A} \in \mathcal{U}} J(\phi, y, \mathcal{A})$, (4.17) derives (4.16). □

For §§ 5 and 6, we will introduce a subsidiary relaxed system. $\mathcal{R} = (W, \mu)$ is called a constant relaxed system, if $\mu'(t, du, \omega) = \nu(du)$ for any t and ω. In this case, we will call μ a constant relaxed control ν and denote $\mathcal{R} = (W, \nu)$. Stressing the terminal time T, we put

$$\begin{aligned} &\mathcal{J}(T, \phi, y, \nu) = J(T, \phi, y, \mu), \quad \text{if } \mu' = \nu \quad (\in \mathcal{P}(\Gamma)) \\ &v(T, \phi, y) = \inf_{\nu \in \mathcal{P}(\Gamma)} \mathcal{J}(T, \phi, y, \nu) \\ &\mathcal{X}(T, \phi, y) = \{\nu \in \mathcal{P}(\Gamma);\ v(T, \phi, y) = \mathcal{J}(T, \phi, y, \nu)\}. \end{aligned} \tag{4.18}$$

Appealing to the fact "$\mathcal{R}_n = (W_n, \nu_n)$ converges to $\mathcal{R} = (W, \nu)$ if and only if $\nu_n \to \nu$ weakly," we get Theorem 4.3.

THEOREM 4.3. *Under the conditions* (A.1)$\sim$(A.3) *and* $(A.4)_{m-2}$, $\mathcal{X}(T, \phi, y)$ *is nonempty and compact. Moreover, there is a Borel selector* $\mathcal{S}_{T,r}$ *of* $\mathcal{X}(T, \phi, y)$, *for* $(\phi, y) \in \Phi_r \times \mathbb{R}^{d'}$.

We consider the following usual pay-off function for the Bellman principle. Let $h: H^{m-2} \times \mathbb{R}^{d'} \to \mathbb{R}^1$ be quadratic growth and satisfy (4.19), namely,

$$|h(\phi, y)| \leqq C(1 + \|\phi\|_{m-2}^2 + |y|^2)$$

and

$$\begin{aligned} &|h(\phi_1, y_1) - h(\phi_2, y_2)| \\ &\quad \leqq C((\|\phi_1\|_{m-2} + \|\phi_2\|_{m-2})\|\phi_1 - \phi_2\|_{m-2} + (|y_1| + |y_2|)|y_1 - y_2|). \end{aligned} \tag{4.19}$$

By $\mathcal{G}$ we denote the set of functions $g: H^m \times \mathbb{R}^{d'} \to \mathbb{R}^1$, which satisfy (4.20) and (4.21) below,

$$|g(\phi, y)| \leqq C_g(1+\|\phi\|_{m-2}^2+|y|^2) \tag{4.20}$$

and

(4.21) for any $\varepsilon, b>0$, there is $\delta=\delta(\varepsilon, b, g)>0$ such that, for $(\phi_i, y_i)\in B_b$ $(=\{(\phi, y)\in H^m\times\mathbb{R}^{d'};\ \|\phi\|_{m-2}<b,\ |y|<b\})$

$$|g(\phi_1, y_1)-g(\phi_2, y_2)|<\varepsilon$$

holds, whenever $\|\phi_1-\phi_2\|_{m-2}<\delta$ and $|y_1-y_2|<\delta$.

Define J and V by (4.22) and (4.23), respectively,

$$J(t, \phi, y, \mathcal{R}, g)=E\left[\int_0^t h(q(s), y+W(s))\,ds+g(q(t), y+W(t))\right] \tag{4.22}$$

where $q=q(\ \cdot\ , \phi, y, \mathcal{R})$, and

$$V(t, \phi, y, g)=\inf_{\mathcal{R}\in\mathfrak{R}} J(t, \phi, y, \mathcal{R}, g). \tag{4.23}$$

For a constant relaxed system, we define $\mathcal{J}$ and v in the same way.

PROPOSITION 4.5. *$J(t, \cdot, \cdot, \mathcal{R}, g)$, $V(t, \cdot, \cdot, g)$, $\mathcal{J}(t, \cdot, \cdot, \mathcal{R}, g)$ and $v(t, \cdot, \cdot, g)$ belong to $\mathcal{G}$, whenever $g\in\mathcal{G}$.*

Proof. From (2.13), we see

$$\begin{aligned}|J(t, \phi, y, \mathcal{R}, g)| &\leqq E\left[C\int_0^t \{1+\|q(s)\|_{m-2}^2+|y+W(s)|^2\}\right.\\ &\qquad \left.+C_g\{1+\|q(t)\|_{m-2}^2+|y+W(t)|^2\}\right] \qquad (4.24)\\ &\leqq C(t, g)(1+\|\phi\|_{m-2}^2+|y|^2)\end{aligned}$$

where $C(t, g)$ is independent of $\mathcal{R}$. So J, V, $\mathcal{J}$, and v also satisfy the quadratic growth condition (4.20).

Recalling Corollary 3.2, we will show (4.21). Put $q_i=q(\ \cdot\ , \phi_i, y_i, \mathcal{R})$ for $(\phi_i, y_i)\in B_b$. Then we have

$$\begin{aligned}&E\left[\int_0^t |h(q_1(s), y_1+W(s))-h(q_2(s), y_2+W(s))|\,ds\right]\\ &\leqq CE\left[\int_0^t \sum_{i=1}^2 \|q_i(s)\|_{m-2}\|q_1(s)-q_2(s)\|_{m-2}+\sum_{i=1}^2 |y_i+W(s)||y_1-y_2|\,ds\right]\\ &\leqq C\left(\sum_{i=1}^2\left\{E\left[\int_0^t \|q_i(s)\|_{m-2}^2\,ds\right]\right\}^{1/2}\left\{E\left[\int_0^t \|q_1(s)-q_2(s)\|_{m-2}^2\,ds\right]\right\}^{1/2}\right. \qquad (4.25)\\ &\qquad \left.+\sum_{i=1}^2 (|y_i|+\sqrt{t})|y_1-y_2|\right)\\ &\leqq C_1(t)(1+b)(|y_1-y_2|(1+b)+\|\phi_1-\phi_2\|_{m-2})\\ &\leqq C_1(t)(1+b)^2(|y_1-y_2|+\|\phi_1-\phi_2\|_{m-2})\end{aligned}$$

$$
\begin{aligned}
&E[|g(q_i(t), y_i + W(t))|; \|q_i(t)\|_{m-2} > n] \\
&\quad \leqq C_g E[1 + \|q_i(t)\|_{m-2}^2 + |y_i + W(t)|^2; \|q_i(t)\|_{m-2} > n] \\
&\quad \leqq C_g (E[1 + \|q_i(t)\|_{m-2}^4 + |y_i + W(t)|^4])^{1/2} (E[\|q_i(t)\|_{m-2}^2 / n^2])^{1/2} \qquad (4.26) \\
&\quad \leqq C_2(g, t)(1 + \|\phi_i\|_{m-2}^2 + |y_i|^2)(1 + \|\phi_i\|_{m-2} + |y_i|)/n \\
&\quad \leqq C_3(g, t)(1+b)^3/n \\
&E[|g(q_i(t), y_i + W(t))|; |y_i + W(t)| > n] \\
&\quad \leqq C_4(g, t)(1 + \|\phi_i\|^2 + |y_i|^2)(|y_i| + \sqrt{t})/n \qquad (4.27) \\
&\quad \leqq C_5(g, t)(1+b)^3/n.
\end{aligned}
$$

Taking a large enough $n = n(\varepsilon, b, t, g)$ such that

$$
(C_3(g, t) + C_5(g, t))(1+b)^3 < \varepsilon n/4, \qquad (4.28)
$$

we get

$$
\begin{aligned}
&|J(t, \phi_1, y_1, \mathcal{R}, g) - J(t, \phi_2, y_2, \mathcal{R}, g)| \\
&\quad \leqq E\left[C \int_0^t \sum_{i=1}^{2} \|q_i(s)\|_{m-2} \|q_1(s) - q_2(s)\|_{m-2} + \sum_{i=1}^{2} |y_i + W(s)| |y_1 - y_2| \, ds \right] \qquad (4.29) \\
&\qquad + E[|g(q_1(t), y_1 + W(t)) - g(q_2(t), y_2 + W(t))|; \\
&\qquad\qquad \|q_i(t)\|_{m-2} < n, \ |y_i + W(t)| < n, \ i = 1, 2] + \varepsilon.
\end{aligned}
$$

From the continuity condition (4.21) for g, we see

$$
\begin{aligned}
&\text{the middle term of the right-hand side of (4.29)} \\
&\quad < \varepsilon + 2C_g(1 + 2n^2) P\{\|q_1(t) - q_2(t)\|_{m-2} > \delta(\varepsilon, n, g)\} \qquad (4.30) \\
&\quad < \varepsilon + 2C_g(1 + 2n^2) E[\|q_1(t) - q_2(t)\|_{m-2}^2 / \delta^2(\varepsilon, n, g)],
\end{aligned}
$$

whenever $|y_1 - y_2| < \delta(\varepsilon, n, g)$.

Using (3.48), (4.29) and (4.30), we can choose a positive constant $\tilde{\delta} = \tilde{\delta}(t, \varepsilon, b, g)$, independent from $\mathcal{R}$, such that

$$
|J(t, \phi_1, y_1, \mathcal{R}, g) - J(t, \phi_2, y_2, \mathcal{R}, g)| < \varepsilon, \qquad (4.31)
$$

whenever $\|\phi_1 - \phi_2\|_{m-2} < \tilde{\delta}$ and $|y_1 - y_2| < \tilde{\delta}$.

Since

$$
\text{“}|V(t, \phi_1, y_1, g) - V(t, \phi_2, y_2, g)| \leqq \sup_{\mathcal{R} \in \mathfrak{R}} |J(t, \phi_1, y_1, \mathcal{R}, g) - J(t, \phi_2, y_2, \mathcal{R}, g)|,\text{”}
$$

we can complete the proof. □

Now, applying arguments similar to (4.6)-(4.9), we get the following theorem.

THEOREM 4.4. *Under the conditions* (A.1)-(A.3) *and* $(A.4)_{m-2}$, *there exists an optimal relaxed system* $\mathcal{R}^*(\phi, y)$, *such that* $\pi(\mathcal{R}^*(\phi, y))$ *is Borel measurable with respect to* $(\phi, y) \in \Phi_r \times \mathbb{R}^{d'}$, *i.e.*,

$$
J(t, \phi, y, \mathcal{R}^*(\phi, y), g) = \inf_{\mathcal{R} \in \mathfrak{R}} J(t, \phi, y, \mathcal{R}, g).
$$

Example. Quadratic loss. (I) Put $h(\phi, y) = \|\phi\|^2$ $(= \|\phi\|_0^2)$ and $g = 0$. Then h satisfies (4.19). So there exists an optimal relaxed system $\mathcal{R}^* = \mathcal{R}^*(\phi, y)$, i.e.,

$$
\min_{\mathcal{R} \in \mathfrak{R}} E\left[\int_0^T \|q(t, \phi, y, \mathcal{R})\|^2 \, dt \right] = E\left[\int_0^T \|q(t, \phi, y, \mathcal{R}^*)\|^2 \, dt \right]
$$

(II) Put $h=0$ and $g(\phi)=\|\phi\|^2$. Then $g\in\mathcal{G}$. So there exists an optimal relaxed system $\tilde{\mathcal{R}}=\tilde{\mathcal{R}}(\phi, y)$, i.e.,

$$\min_{\mathcal{R}\in\mathfrak{R}} E[\|q(T, \phi, y, \mathcal{R})\|^2]=E[\|q(T, \phi, y, \tilde{\mathcal{R}})\|^2].$$

5. Approximation. In this section, we will show that there exists an approximate optimal control which is adapted to a Wiener process.

We call $\mathcal{R}=(W, \mu)$ a step relaxed system, if $\mu'(t)=\mu'([t/\Delta]\Delta)$ with a positive Δ, where $[\ ]=$Gauss symbol. By $\mathfrak{R}_N$ we denote the totality of step relaxed systems with $\Delta=2^{-N}$. For μ we define an approximate derivative μ'_n as follows:

$$\mu'_n(t, \cdot)=\begin{cases}2^n\mu([t-2^{-n}, t)\times\cdot) & \text{for } t>2^{-n}\\ t^{-1}\mu([0, t)\times\cdot) & \text{for } t\leqq 2^{-n}.\end{cases} \tag{5.1}$$

Put

$$\mu'_{n,k}(t, \cdot)=\mu'_n([2^k t]2^{-k}, \cdot) \tag{5.2}$$

and $\mu_{n,k}([0, t]\times A)=\int_0^t \mu'_{n,k}(s, A)\,ds$. Then, for a suitable sequence $k(n)$, $n=1, 2, \cdots$, we have, with probability 1,

$$\mu_{n,k(n)}\to\mu \text{ weakly.} \tag{5.3}$$

Hereafter, we consider a pay-off function J as (4.22). Therefore, (5.3) yields

$$V(t, \phi, y, g)=\lim_{N\to\infty}\inf_{\mathfrak{R}_N} J(t, \phi, y, \mathcal{R}, g). \tag{5.4}$$

Putting

$$\tilde{\mathfrak{R}}_N=\{\mathcal{R}=(W, \mu)\in\mathfrak{R}_N;\ \mu \text{ is } W\text{-adapted}\}, \tag{5.5}$$

we have the following theorem.

THEOREM 5.1. *Under the conditions* (A.1)-(A.3) *and* $(A.4)_{m-2}$, *we have, for* $\phi\in\Phi$

$$\inf_{\mathfrak{R}_N} J(t, \phi, y, \mathcal{R}, g)=\inf_{\tilde{\mathfrak{R}}_N} J(t, \phi, y, \mathcal{R}, g). \tag{5.6}$$

Proof. Since $\tilde{\mathfrak{R}}_N\subset\mathfrak{R}_N$, it is enough to show

$$J(t, \phi, y, \mathcal{R}, g)\geqq\inf_{\tilde{\mathfrak{R}}_N} J(t, \phi, y, \mathcal{R}, g) \quad \forall\mathcal{R}\in\mathfrak{R}_N. \tag{5.7}$$

Putting $\Delta=2^{-N}$ and $j\Delta<t\leqq(j+1)\Delta$, we will evaluate I, defined by (5.8),

$$I=E\left[\int_{j\Delta}^t h(q(s), y+W(s))\,ds+g(q(t), y+W(t))\,|\,\mathcal{F}_{j\Delta}\right] \tag{5.8}$$

where $q=q(\ , \phi, y, \mathcal{R})$. Under the conditional probability $P(\cdot\,|\,\mathcal{F}_{j\Delta})$, $W^j(\cdot)=W(\cdot+j\Delta)-W(j\Delta)$ becomes a new Wiener process which is independent of $\mathcal{F}_{j\Delta}$ and $\mu'(\theta+j\Delta, \cdot)=\mu'(j\Delta, \cdot)$, $0\leqq\theta\leqq t-j\Delta$, can be regarded as a constant relaxed control. Moreover, the uniqueness of solution derives

$$q(\theta+j\Delta, \phi, y, \mathcal{R})=q(\theta, q(j\Delta, \phi, y, \mathcal{R}), y+W(j\Delta), \mu'(j\Delta)) \quad \text{for } 0\leqq\theta\leqq t-j\Delta. \tag{5.9}$$

Hence, we see

$$\begin{aligned} I&\geqq\inf_{\nu\in\mathcal{P}(\Gamma)}\mathcal{J}(t-j\Delta, q(j\Delta, \phi, y, \mathcal{R}), y+W(j\Delta), \nu, g)\\ &=v(t-j\Delta, q(j\Delta, \phi, y, \mathcal{R}), y+W(j\Delta), g).\end{aligned} \tag{5.10}$$

Defining $v(s): \mathcal{G} \to \mathcal{G}$ by $v(s, \cdot, g) = v(s)g$, we see from (5.10)

$$J(t, \phi, y, \mathcal{R}, g) \geqq E\left[\int_0^{j\Delta} h(q(s), y + W(s))\, ds + v(t - j\Delta)g(q(j\Delta), y + W(j\Delta))\right]. \tag{5.11}$$

By the same argument, we calculate $E[\cdots | \mathcal{F}_{(j-1)\Delta}]$ and obtain

$$J(t, \phi, y, \mathcal{R}, g) \geqq E\left[\int_0^{(j-1)\Delta} h(q(s), y + W(s))\, ds + v(\Delta)v(t - j\Delta)g(q((j-1)\Delta), y + W((j-1)\Delta))\right]. \tag{5.12}$$

Repeating this evaluation, we get

$$J(t, \phi, y, \mathcal{R}, g) \geqq v^j(\Delta)v(t - j\Delta)g(\phi, y). \tag{5.13}$$

We assume that $(A.4)_{m-2, r_0}$ holds. Then (2.14) asserts that $q(t, \phi, y, \mathcal{R}) \in \Phi_r$ with probability 1, whenever $\phi \in \Phi_r$ for $r \leqq r_0$. According to Theorem 4.3, we can take a Borel selector $\mathcal{S}_r(t, g)$ of

$$\mathfrak{X}(t, \phi, y, g) = \{\pi(\mathcal{R}); v(t, \phi, y, g) = J(t, \phi, y, \mathcal{R}, g)\}.$$

Let $(\Omega_i, \mathcal{F}_i, P_i)$, $i = 1, \cdots, (j+1)$ be a probability space and W_i be a Wiener process on it. Let us set

$$\Omega = \prod_{i=1}^{j+1} \Omega_i, \quad \mathcal{F} = \prod_{i=1}^{j+1} \mathcal{F}_i, \quad P = \prod_{i=1}^{j+1} P_i$$

and

$$W(t) = \begin{cases} W_1(t) & \text{for } 0 \leqq t < \Delta \\ W_1(\Delta) + W_2(t - \Delta) & \text{for } \Delta \leqq t < 2\Delta \\ \vdots & \\ \sum_{k=1}^{j} W_k(\Delta) + W_{j+1}(t - j\Delta) & \text{for } j\Delta \leqq t < (j+1)\Delta. \end{cases} \tag{5.14}$$

Then W becomes a Wiener process on $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, where $\mathcal{F}_t = \sigma_t(W)$. Fix $\nu_1 \in \mathfrak{X}(\Delta, \phi, y, v^{j-1}(\Delta)v(t - j\Delta)g)$ arbitrarily and q_1 denotes the solution of (5.15).

$$\begin{aligned} &dq_1(t) = \tilde{L}(y + W(t), \nu_1)q_1(t)\, dt + M(y + W(t))q_1(t)\, dW(t), \\ &q_1(0) = \phi \qquad 0 < t \leqq \Delta. \end{aligned} \tag{5.15}$$

So q_1 is W-adapted.

Put $\nu_2 = \mathcal{S}_r(\Delta, v^{j-2}(\Delta)v(t - j\Delta)g(q_1(\Delta), y + W(\Delta))$ and q_2 denotes the solution of (5.16).

$$\begin{aligned} &dq_2(t) = \tilde{L}(y + W(t), \nu_2)q_2(t)\, dt + M(y + W(t))q_2(t)\, dW(t), \\ &q_2(\Delta) = q_1(\Delta) \qquad \Delta < t \leqq 2\Delta. \end{aligned} \tag{5.16}$$

Putting $\nu_3 = \mathcal{S}_r(\Delta, v^{j-2}(\Delta)v(t - j\Delta)g(q_2(2\Delta), y + W(2\Delta))$, we repeat the same argument. Now define μ' by

$$\mu'(t) = \nu_k \quad \text{for } t \in [(k-1)\Delta, k\Delta). \tag{5.17}$$

Then μ' is W-adapted and $\tilde{\mathcal{R}} = (\Omega, \mathcal{F}, \mathcal{F}_t, P, W, \mu) \in \tilde{\mathfrak{R}}_N$. Moreover, putting $q = q_k$ on $[(k-1)\Delta, k\Delta)$, we get

$$
\begin{aligned}
&E\left(\int_{j\Delta}^{t} h(q(s), y+W(s))\,ds + g(q(t), y+W(t)) \,\middle|\, \mathcal{F}_{j\Delta}\right) \\
&\quad = E\left(\int_{j\Delta}^{t} h(q_j(s), y+W(s))\,ds + g(q_j(t), y+W(t))\right) \\
&\quad = v(t-j\Delta)g(q(j\Delta), y+W(j\Delta)), \\
&E\left(\int_{(j-1)\Delta}^{j\Delta} h(q(s), y+W(s))\,ds + v(t-j\Delta)g(q(j\Delta), y+W(j\Delta)) \,\middle|\, \mathcal{F}_{(j-1)\Delta}\right) \\
&\quad = v(\Delta)v(t-j\Delta)g(q((j-1)\Delta), y+W((j-1)\Delta)),
\end{aligned} \tag{5.18}
$$

and so on. Thus, we have

$$
\begin{aligned}
&J(t, \phi, y, \tilde{\mathcal{R}}, g) \\
&\quad = E\left[\int_0^t h(q(s), y+W(s))\,ds + g(q(t), y+W(t))\right] \\
&\quad = E\left[E\left(\int_{j\Delta}^{t} h(q(s), y+W(s))\,ds + g(q(t), y+W(t)) \,\middle|\, \mathcal{F}_{j\Delta}\right)\right. \\
&\qquad\qquad \left. + \int_0^{j\Delta} h(q(s), y+W(s))\,ds\right] \\
&\quad = E\left[\int_0^{j\Delta} h(q(s), y+W(s))\,ds + v(t-j\Delta)g(q(j\Delta), y+W(j\Delta))\right] \\
&\quad = E\left[\int_0^{(j-1)\Delta} h(q(s), y+W(s))\,ds\right. \\
&\qquad\qquad \left. + v(\Delta)v(t-j\Delta)g(q((j-1)\Delta), y+W((j-1)\Delta))\right] \\
&\quad = v^j(\Delta)v(t-j\Delta)g(\phi, y).
\end{aligned} \tag{5.19}
$$

From (5.13) and (5.19), we can conclude (5.7). □

Recalling (5.4), we obtain Corollary 5.1.

COROLLARY 5.1. *Under the same condition of Theorem* 5.1,

$$V(t, \phi, y, g) = \lim_{N\to\infty} \inf_{\mathfrak{R}_N} J(t, \phi, y, \mathcal{R}, g) \tag{5.20}$$

holds. In other words, there is an approximate optimal step relaxed system, which is adapted to a Wiener process.

Using the chattering lemma [3], $\mathcal{R} \in \mathfrak{R}_N$ can be approximated by admissible controls which are adapted to a Wiener process. Hence, putting $\tilde{\mathfrak{U}}_N = \mathfrak{U} \cap \tilde{\mathfrak{R}}_N = \{\mathcal{A} = (W, U);\ U \text{ is } W\text{-adapted and } U(t) = U([2^N t]2^{-N})\}$ and $\tilde{\mathfrak{U}} = \bigcup_{N=1}^{\infty} \tilde{\mathfrak{U}}_N$, we have Corollary 5.2.

COROLLARY 5.2. *Under the same condition, there is an approximate optimal step system* $\mathcal{A} \in \tilde{\mathfrak{U}}$.

6. Bellman principle. Now we are ready to prove the Bellman principle. For $\phi \in \Phi_r$ and $\mathscr{A} = (W, U) \in \tilde{\mathfrak{U}}_N$, we will evaluate (6.1):

$$
\begin{aligned}
&J(s+t, \phi, y, \mathscr{A}, g) \\
&\quad = E\left(\int_0^t h(q(\theta), y + W(\theta))\, d\theta \right. \\
&\qquad \left. + E\left[\int_t^{t+s} h(q(\theta), y + W(\theta))\, d\theta + g(q(t+s), y + W(t+s)) \,|\, \mathscr{F}_t \right] \right).
\end{aligned} \tag{6.1}
$$

Since $W^t(\cdot) = W(\cdot + t) - W(t)$ is a Wiener process independent from $\mathscr{F}_t$, we see

(6.2) conditional expectation of 2nd term $\geqq V(s, q(t), y + W(t), g)$ w.p. 1.

This asserts

$$
\begin{aligned}
J(s+t, \phi, y, \mathscr{A}, g) &\geqq J(t, q(t), y + W(t), \mathscr{A}, V(s, \cdot, g)) \\
&\geqq V(t, q(t), y + W(t), V(s, \cdot, g)).
\end{aligned} \tag{6.3}
$$

Now Corollary 5.2 yields

$$
V(s+t, \phi, y, g) \geqq V(t, q(t), y + W(t), V(s, \cdot, g)). \tag{6.4}
$$

Next we will show the converse inequality of (6.4) by a standard argument.

Let $\mathscr{S}_1(\phi, y)$ denote a Borel selector of

$$
\mathfrak{X}(\phi, y) = \{\pi(\mathscr{R});\ V(s, \phi, y, g) = J(s, \phi, y, \mathscr{R}, g)\}.
$$

For any $\mathscr{A} = (\Omega, \mathscr{F}, \mathscr{F}_t, P, W, U) \in \tilde{\mathfrak{U}}_N$, we put $\tilde{\Omega} = C([0, s]; \mathbb{R}^{d'}) \times \Lambda$, $\tilde{W}$ = first coordinate function $\tilde{\mu}$ = second coordinate function, $\tilde{\mathscr{F}} = \sigma(\tilde{W}, \tilde{\mu})$, $\tilde{\mathscr{F}}_\theta = \sigma_\theta(\tilde{W}, \tilde{\mu})$ $\Omega^* = \Omega \times \tilde{\Omega}$, $\mathscr{F}^* = \mathscr{F} \times \tilde{\mathscr{F}}$.

Define P^* by

$$
P^*((\tilde{W}, \tilde{\mu}) \in B \,|\, \mathscr{F}_t) = \mathscr{S}_r(q(t, \phi, y, \mathscr{A}), y + W(t))(B), \tag{6.5}
$$

namely,

$$
P^*((\tilde{W}, \tilde{\mu}) \in B, (W, \mu) \in C) = \int_{\{(W, \mu) \in C\}} \mathscr{S}_r(q(t, \phi, y, \mathscr{A}), y + W(t))(B)\, dP.
$$

Hence, $\tilde{W}$ is a Wiener process on $(\Omega^*, \mathscr{F}^*, P^*)$, independent from W. Thus, putting

$$
W^*(\theta) = \begin{cases} W(\theta), & \theta \leqq t \\ W(t) + \tilde{W}(\theta - t), & t \leqq \theta \leqq s + t \end{cases}
$$

$$
\mu^*(\theta, \cdot) = \begin{cases} \delta_{U(\theta)}(\cdot), & \theta \leqq t \\ \tilde{\mu}(\theta - t, \cdot), & t \leqq \theta \leqq s + t \end{cases}
$$

$$
\mathscr{F}^*_\theta = \sigma_\theta(W^*, \mu^*),
$$

we see $\mathscr{R}^* = (W^*, \mu^*) \in \mathfrak{R}$ and its response q^* satisfies

$$
\begin{aligned}
&E\left[\int_t^{t+s} h(q^*(\theta), y + W^*(\theta))\, d\theta + g(q^*(t+s), y + W(t+s)) \,|\, \mathscr{F}^*_t \right] \\
&\quad = V(s, q(t, \phi, y, \mathscr{A}), y + W(t), g) \\
&\quad = V(s, q(t, \phi, y, \mathscr{R}^*), y + W^*(t), g).
\end{aligned} \tag{6.6}
$$

Therefore,

$$J(s+t, \phi, y, \mathcal{R}^*, g) = J(t, \phi, y, \mathcal{R}^*, V(s, \cdot, g)) \tag{6.7}$$
$$= J(t, \phi, y, \mathcal{A}, V(s, \cdot, g))$$

holds. This asserts

$$V(s+t, \phi, y, g) \leqq J(t, \phi, y, \mathcal{A}, V(s, \cdot, g)).$$

Again, Corollary 5.2 concludes the converse inequality of (6.4).

Thus, we obtain Theorem 6.1.

THEOREM 6.1. *Under the conditions* (A.1)-(A.3) *and* $(A.4)_{m-2}$, *we have*

$$V(t, \cdot, g) \in \mathcal{G} \quad \textit{whenever } g \in \mathcal{G}, \tag{6.8}$$

and the Bellman principle holds, i.e.,

$$V(s+t, \phi, y, g) = V(t, \phi, y, V(s, \cdot, g)) \quad \textit{for } \phi \in \Phi \quad \textit{and} \quad g \in \mathcal{G}. \tag{6.9}$$

Remark. The Bellman principle is formulated by some nonlinear group [1].

7. Applications. (1) Temperature control. Let us consider a heat system in a random medium. The field of temperature $q(t, x)$ is governed by the following SPDE.

$$dq(t, x) = (\Delta q(t, x) + f(x, U(t)))\, dt + g(x)\, dW(t), \qquad t > 0, \quad x \in \mathbb{R}^d,$$

with the initial data $q(0, x) = \phi(x)$, where Δ is the Laplacian operator for x and W a d-dimensional Wiener process. So the temperature is controlled through the external force $f(x, U(t))$. The problem is to minimize the deviation of temperature distribution from the assigned distribution m at a given time T (cf. Sakawa [18]), namely, the pay-off function J is defined by

$$J(U) = E\left[\int_{\mathbb{R}^d} |q(T, x) - m(x)|^2\, dx\right].$$

Hence, Theorem 4.4 concludes the existence of an optimal relaxed control, if f and g satisfy the condition $(A.4)_1$.

(2) Nervous system. In Chapter 3 of [20], Walsh deals with the following SPDE as the dynamics of nervous system,

$$dq(t, x) = \left(\frac{\partial^2}{\partial x^2} q(t, x) - q(t, x)\right) dt + (q(t, x) - g(x))\, dW(t), \qquad 0 < x < L, \quad t > 0,$$

with Neuman boundary condition, where W is a one-dimensional Wiener process, and also considers the barrier problem.

Since a medical treatment acts as an external force, here we will consider the following SPDE as its variant,

$$dq(t, x) = \left(\frac{\partial^2}{\partial x^2} q(t, x) - q(t, x) + f(x, U(t))\right) dt$$
$$+ (q(t, x) - g(x))\, dW(t), \qquad x \in \mathbb{R}^1, \qquad 0 < t \leqq T,$$
$$q(0, x) = \phi(x).$$

Although we want to keep $q(t, x)$ near an assigned level λ at a given spot y, we need some smooth modifications. For given two positive constants b and c, we put

$$p(t) = \frac{1}{2c} \int_{-c}^{c} q(t, y + x)\, dx$$

and

$$h(x)=\begin{cases}1, & x\notin(\lambda-b,\lambda+b)\\ \dfrac{\lambda-x}{b}, & \lambda-b<x<\lambda\\ \dfrac{x-\lambda}{b}, & \lambda<x<\lambda+b.\end{cases}$$

Now the problem is to minimize $E[\int_0^T h(p(t))\,dt]$ and our theorems are applicable.

(3) Stochastic control with partial observation.

Let B and $\hat{B}$ be independent Wiener processes with values in $\mathbb{R}^{d'}$ and $\mathbb{R}^d$, respectively. Suppose that the d-dimensional state process X and the d'-dimensional observation process Y are governed by the following stochastic differential equations (SDE in short) with bounded and smooth coefficients:

$$\begin{aligned}dX(t)&=\gamma(X(t),Y(t),U(t))\,dt+\alpha(X(t),Y(t),U(t))\,d\hat{B}(t)\\&\quad+b(X(t),Y(t))\,dB(t),\qquad 0<t\leqq T\\X(0)&=\xi\end{aligned}\tag{7.1}$$

and

$$dY(t)=f(X(t))\,dt+dB(t),\qquad Y(0)=0,\tag{7.2}$$

where U is an admissible control. So in our model the state and observation noises may not be independent.

Let h and $G:\mathbb{R}^d\times\mathbb{R}^{d'}\to\mathbb{R}^1$, be bounded and Lipschitz continuous. The problem is to minimize the pay-off function J, defined by

$$J(U)=E\left[\int_0^T h(X(t),Y(t))\,dt+G(X(T),Y(T))\right]\tag{7.3}$$

by a suitable choice of U.

In the customary version of stochastic control with partial observation, $U(t)$ is a function of the observation process $Y(s)$, $s\leqq t$, namely, admissible control in the strict sense. Here we treat some wider class of admissible controls, according to [4], as follows:

$A=(\Omega,\mathcal{F},\mathcal{F}_t,\mathring{P},\hat{B},Y,U)$ is called an admissible control system, if

(I) $(\Omega,\mathcal{F},\mathcal{F}_t,\mathring{P})$ is a probability space, with $\mathcal{F}_t=\sigma_t(Y,U)$

(II) Y is a d'-dimensional $\mathcal{F}_t$-Wiener process

(III) U is a Γ-valued process

(IV) $\hat{B}$ is a d-dimensional Wiener process on Ω, independent from (Y,U).

Let ξ be a random variable independent from $(\hat{B},Y,U)$ and ϕ be its probability density. For an admissible system A, we consider SDE,

$$\begin{aligned}dX(t)&=(\gamma(X(t),Y(t),U(t))-b(X(t),Y(t))f(X(t)))\,dt\\&\quad+\alpha(X(t),Y(t),U(t))\,d\hat{B}(t)+b(X(t),Y(t))\,dY(t)\\X(0)&=\xi.\end{aligned}\tag{7.4}$$

Put

$$\rho(t)=\exp\left\{\int_0^t f(X(t))\,dY(t)-\frac{1}{2}\int_0^t|f(X(t))|^2\,dt\right\}\tag{7.5}$$

and define a new probability P by

$$dP = \rho(T)\, d\mathring{P}. \tag{7.6}$$

Then, Girsanov's theorem asserts that, under the probability P, $B(t) = Y(t) - \int_0^t f(X(s))\, ds$, $0 \leqq t \leqq T$, turns out to be a Wiener process independent from $\hat{B}$, and (X, Y) satisfies (7.1).

Moreover, the pay-off function $J(U)$ of (7.3) can be written by

$$J(A) = \mathring{E}\left[\int_0^T h(X(t), Y(t))\rho(t)\, dt + G(X(T), Y(T))\rho(T)\right]$$

where $\mathring{E}$ means the expectation with respect to $\mathring{P}$.

On the other hand, $A = (\Omega, \mathscr{F}, \mathscr{F}_t, \mathring{P}, \hat{B}, Y, U)$ derives an admissible system $\mathscr{A} = (\Omega, \mathscr{F}, \mathscr{F}_t, \mathring{P}, Y, U)$, and an admissible system turns out to be an admissible control system when we add an independent Wiener process $\hat{B}$. For $\mathscr{A} = (\Omega, \mathscr{F}, \mathscr{F}_t, \mathring{P}, Y, U)$, we consider SPDE,

$$\begin{aligned} &dq(t) = L(Y(t), U(t))q(t)\, dt + M(Y(t))q(t)\, dY(t) \\ &q(0) = \phi \ (\in H^3) \end{aligned} \tag{7.7}$$

where

$$\begin{aligned} &L(y, u)q = \sum_{i,j=1}^{d} \frac{\partial}{\partial x_j} a_{ij}(\cdot, y, u) \frac{\partial}{\partial x_i} q - \sum_{j=1}^{d} \frac{\partial}{\partial x_j} (\tilde{a}_j(\cdot, y, u)q), \\ &M^k(y)q = -\sum_{i=1}^{d} b_{ik}(\cdot, y) \frac{\partial}{\partial x_i} q + \tilde{f}_k(\cdot, y)q, \\ &a(x, y, u) = (b(x, y)b^*(x, y) + \alpha(x, y, u)\alpha^*(x, y, u))/2, \\ &\tilde{a}_j(x, y, u) = \gamma_j(x, y, u) - \sum_{i=1}^{d} \frac{\partial a_{ij}}{\partial x_i}(x, y, u), \end{aligned} \tag{7.8}$$

and

$$\tilde{f}_k(x, y) = f_k(x) - \sum_{i=1}^{d} \frac{\partial b_{ik}}{\partial x_i}(x, y).$$

Then, under the conditions (A.1)-(A.3), $J(A)$ can be represented by

$$J(A) = \mathring{E}\left[\int_0^T (h(\cdot, Y(t)), q(t))\, dt + (G(\cdot, Y(T)), q(T))\right]. \tag{7.9}$$

Now we have the following theorem, appealing to Theorems 3.1 and 4.2.

THEOREM 7.1. *Suppose* (A.1)-(A.3), $(A.4)_1$ *and the convexity condition for the coefficients of the* SPDE (7.7). *Then, for* $\phi \in \Phi$, *there is an optimal admissible control system* A^*, *namely*

$$J(A^*) = \inf_A J(A). \tag{7.10}$$

Appendix. Let us prove the lemma in § 3. Here we use the following notation, according to [10]:

For $\alpha = (i_1, \cdots, i_l)$, $D^\alpha = \partial_{i_1} \cdots \partial_{i_l}$, $|\alpha| = l$

$\binom{\alpha}{\gamma} = \binom{i_1}{j_1} \cdots \binom{i_l}{j_l}$ is the binomial coefficient

(for $\gamma = (j_1, \cdots, j_l)$, $0 \leqq j_k \leqq i_k$)

$|i| = 0$ for $i = 0$, $= 1$ for $i = 1, \cdots, d$,

$\int \cdots dx$ stands for $\int_{\mathbb{R}^d} \cdots dx$ and hereafter $N_1, N_2, \cdots$ denote constants depending only on K, T, and l, and repeated indexes are assumed to be summed from 1 (not 0) to d.

We will estimate the principal part of J defined by (1). For $u \in C_0^\infty(\mathbb{R}^d)$, we put

$$
\begin{aligned}
J &= \sum_{|\gamma| \leqq l} \int \{-2D^\gamma(a^{ij}\partial_j u)D^\gamma \partial_i u + 3D^\gamma(b^i \partial_i u)D^\gamma(b^j \partial_j u)\}\, dx \\
&= -2 \int \hat{a}^{ij} \partial_i u \partial_j u\, dx \qquad (1) \\
&\quad + \sum_{1 \leqq |\gamma| \leqq l} \int \{-2D^\gamma(a^{ij}\partial_j u)D^\gamma \partial_i u + 3D^\gamma(b^i \partial_i u)D^\gamma(b^j \partial_j u)\}\, dx,
\end{aligned}
$$

where $\hat{a}^{ij} = a^{ij} - \frac{3}{2} b^i \cdot b^j$.

Using integration by parts, we get

$$
\begin{aligned}
&\int -2D^\gamma(a^{ij}\partial_j u)D^\gamma \partial_i u\, dx \\
&\quad = \int -2a^{ij} D^\gamma \partial_j u D^\gamma \partial_i u\, dx + 2 \sum_{\substack{\alpha+\beta=\gamma \\ |\alpha|=1}} \binom{\gamma}{\alpha} \int D^\alpha a^{ij} D^\beta \partial_i \partial_j u D^\gamma u\, dx \\
&\quad\quad + 2 \sum_{\substack{\alpha+\beta=\gamma \\ |\alpha| \geqq 1}} \binom{\gamma}{\alpha} \int D^\alpha \partial_i a^{ij} D^\beta \partial_j u D^\gamma u\, dx \qquad (2) \\
&\quad\quad + 2 \sum_{\substack{\alpha+\beta=\gamma \\ |\alpha| \geqq 2}} \binom{\gamma}{\alpha} \int D^\alpha a^{ij} D^\beta \partial_i \partial_j u D^\gamma u\, dx.
\end{aligned}
$$

Appealing to $|\beta|+1 \leqq l$ in the third term and $|\beta|+2 \leqq l$ in the fourth term,

$$\leqq \text{first term} + \text{second term} + N_1 \|u\|_l^2.$$

Since $D^\gamma(b^i \partial_i u) - b^i D^\gamma \partial_i u$ is independent of the $(l+1)$th order derivative of u, we obtain, in the same way as (2),

$$
\begin{aligned}
&\int 3D^\gamma(b^i \partial_i u) \cdot D^\gamma(b^j \partial_j u)\, dx \\
&\quad = \int 3|b^i D^\gamma \partial_i u + (D^\gamma(b^i \partial_i u) - b^i D^\gamma \partial_i u)|^2\, dx \\
&\quad = \int 3b^i D^\gamma \partial_i u \cdot b^j D^\gamma \partial_j u\, dx + \int 6 \sum_{\substack{\alpha+\beta=\gamma \\ |\alpha| \geqq 1}} \binom{\gamma}{\alpha} D^\alpha b^i D^\beta \partial_i u \cdot b^j D^\gamma \partial_j u\, dx \\
&\quad\quad + \int 3|D^\gamma(b^i \partial_i u) - b^i D^\gamma \partial_i u|^2\, dx \qquad (3) \\
&\quad \leqq \text{first term} - 6 \sum_{\substack{\alpha+\beta=\gamma \\ |\alpha|=1}} \binom{\gamma}{\alpha} \int (D^\alpha b^i) \cdot b^j D^\beta \partial_i \partial_j u D^\gamma\, dx + N_2 \|u\|_l^2 \\
&\quad = \text{first term} - 3 \sum_{\substack{\alpha+\beta=\gamma \\ |\alpha|=1}} \binom{\gamma}{\alpha} \int D^\alpha (b^i \cdot b^j) D^\beta \partial_i \partial_j u D^\gamma u\, dx + N_2 \|u\|_l^2.
\end{aligned}
$$

(1), (2), and (3) yield

$$J \leqq -2 \sum_{|\gamma| \leqq l} \int \hat{a}^{ij} D^{\gamma} \partial_j u D^{\gamma} \partial_i u \, dx$$

$$+2 \sum_{1 \leqq |\gamma| \leqq l} \sum_{\substack{\alpha+\beta=\gamma \\ |\alpha|=1}} \binom{\gamma}{\alpha} \int D^{\alpha} \hat{a}^{ij} D^{\beta} \partial_i \partial_j u D^{\gamma} u \, dx + N_3 \|u\|_l^2.$$

On the other hand,

$$|D^{\alpha} \hat{a}^{ij} D^{\beta} \partial_i \partial_j u|^2 \leqq N_4 \hat{a}^{ij} D^{\beta} \partial_i \partial_k u D^{\beta} \partial_j \partial_k u$$

$$\leqq N_4 \sum_{|\gamma| \leqq l} \hat{a}^{ij} D^{\gamma} \partial_i u D^{\gamma} \partial_j u$$

holds, by virtue of $\hat{a}^{ij} \in C^2(\mathbb{R}^d)$ and matrix $(\hat{a}^{ij}) \geqq 0$, (see Lemma 1.7.1 of [15]).

Noting $2|ab| \leqq \varepsilon^2 |a|^2 + |b|/\varepsilon^2$, we get

$$J \leqq (N_5 \varepsilon^2 - 2) \sum_{|\gamma| \leqq l} \int \hat{a}^{ij} D^{\gamma} \partial_j u D^{\gamma} \partial_i u \, dx + (N_3 + N_6/\varepsilon^2) \|u\|_l^2.$$

So $J \leqq N_7 \|u\|_l^2$ holds, putting $\varepsilon^2 = 2/N_5$. Applying the same calculation to the other terms, we can prove the lemma for $u \in C_0^{\infty}(\mathbb{R}^d)$. Since $C_0^{\infty}(\mathbb{R}^d)$ is dense in H^{l+1}, we can conclude the proof of the lemma by the routine method. □

REFERENCES

[1] A. BENSOUSSAN AND M. NISIO, *Nonlinear semi-group arising in the control of diffusions with partial observation*, Stochastics, to appear.

[2] N. EL KAROUI, D. HUÙ NGUYEN, AND M. JEANBLANC-PICQUÉ, *Existence of an optimal Markovian filter for the control under partial observations*, SIAM J. Control Optim., 26 (1988), pp. 1025–1061.

[3] W. H. FLEMING AND M. NISIO, *On stochastic relaxed control for partially observed diffusions*, Nagoya Math. J., 93 (1984), pp. 71–108.

[4] W. H. FLEMING AND E. PARDOUX, *Optimal control for partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–285.

[5] Y. FUJITA, *Linear stochastic partial differential equations with constant coefficients*, J. Math. Kyoto Univ., 28 (1988), pp. 301–310.

[6] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, Kodansha/North-Holland, Tokyo, Amsterdam, 1981.

[7] K. ITÔ, *Foundations of Stochastic Differential Equations in Infinite Dimensional Spaces*, CBMS-NSF Regional Conference Series, 47 Society for Industrial and Applied Mathematics, Philadelphia, PA, 1984.

[8] N. V. KRYLOV AND B. L. ROZOVSKII, *On the Cauchy problem for linear stochastic partial differential equations*, Izv. Akad. Nauk SSSR Ser. Mat., 41 (1977), pp. 1329–1347; Math. USSR-Izv., 11 (1977), pp. 1267–1284.

[9] ———, *On conditional distributions of diffusion processes*, Izv. Akad. Nauk SSSR Ser. Mat., 42 (1978), pp. 356–378; Math. USSR-Izv., 12 (1978), pp. 336–356.

[10] ———, *On characteristics of the degenerate parabolic Itô equations of the second order*, Petrovskii Seminar, vol. 8, Moskovskogo Universiteta, 1982, pp. 153–168. (In Russian).

[11] ———, *Stochastic partial differential equations and diffusion processes*, Uspekhi Mat. Nauk, 37 (1982), pp. 75–95; Russian Math. Surveys, 37 (1982), pp. 81–105.

[12] H. KUNITA, *Stochastic partial differential equations connected with non-linear filtering*, Proceedings Cortona, 1981: Lecture Notes in Math. 972, Third session C.I.M.E., Springer-Verlag, New York, Berlin, 1982, pp. 100–169.

[13] J. L. LIONS, *Equations Différentielles Opérationnelles et problèmes aux Limites*, Springer-Verlag, Berlin, 1961.

[14] N. NAGASE, *On the existence of optimal control for controlled stochastic partial differential equations*, Nagoya Math. J., to appear.

[15] O. A. OLEINIK AND E. V. RADKEVICH, *Second Order Equations with Non-negative Characteristic Form*, Plenum Press, New York, 1973.

[16] E. PARDOUX, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics, 3 (1979), pp. 127–167.

[17] B. L. ROZOVSKII, *Nonnegative L^1-solutions of second order stochastic parabolic equations with random coefficients*, Steklov Seminar, Steklov Institute of Mathematics, 1984, Stat. and Control of Stoch. Proc., Trans. Math. Eng., 1985, pp. 410–427.

[18] Y. SAKAWA, *Solution of an optimal control problem in a distributed parameter system*, IEEE Trans. Automat. Control, AC-9 (1964), pp. 420–426.

[19] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, 1979.

[20] J. B. WALSH, *An introduction to stochastic partial differential equations*, Lecture Notes in Math. 1180, Springer-Verlag, New York, Berlin, 1986, pp. 266–437.

[21] QING ZHANG, *Controlled partially observed diffusions with correlated noise*, Appl. Math. Optim., to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 214-242, January 1990

011

DUAL ASCENT METHODS FOR PROBLEMS WITH STRICTLY CONVEX COSTS AND LINEAR CONSTRAINTS: A UNIFIED APPROACH*

PAUL TSENG†

Abstract. Consider problems of the form

$$\text{(P)} \qquad \min \{f(x) \mid Ex \geqq b\},$$

where f is a strictly convex (possibly nondifferentiable) function and E and b are, respectively, a matrix and a vector. A popular method for solving special cases of (P) (e.g., network flow, entropy maximization, quadratic program) is to dualize the constraints $Ex \geqq b$ to obtain a differentiable maximization problem and then apply an iterative ascent method to solve it. This method is simple and can exploit sparsity, thus making it ideal for large-scale optimization and, in certain cases, for parallel computation. Despite its simplicity, however, convergence of this method has been shown only under certain very restrictive conditions and only for certain special cases of (P). In this paper a block coordinate ascent method is presented for solving (P) that contains as special cases both dual coordinate ascent methods and dual gradient methods. It is shown, under certain mild assumptions on f and (P), that this method converges. Also the line searches are allowed to be inexact and, when f is separable, can be done in parallel.

Key words. block coordinate ascent, strict convexity, convex program

AMS(MOS) subject classifications. 49, 90

1. Introduction.

Consider the problem

$$\text{(P)} \qquad \text{Minimize} \quad f(x)$$

$$\text{(1.1)} \qquad \text{subject to} \quad Ex \geqq b,$$

where $f: \Re^m \to (-\infty, +\infty]$, E is an $n \times m$ matrix having no zero row, and b is a vector in $\Re^n$. In our notation all vectors are column vectors and superscript T denotes transpose. We denote by e_{ij} the (i,j)th entry of E and by b_i the ith coordinate of b. We also denote by $\langle \cdot, \cdot \rangle$ the usual Euclidean inner product and by $\| \cdot \|$ its induced norm. On occasion, we will treat in parallel the equality constraint problem

$$(\text{P}^{\text{E}}) \qquad \text{Minimize} \quad f(x)$$

$$\text{subject to} \quad Ex = b.$$

(Extension to the case of mixed equality and inequality constraints is straightforward.)

Denote by S the *effective domain* of f, i.e.,

$$S = \{x \in \Re^m \mid f(x) < +\infty\},$$

by int (S), ri (S) and cl (S), respectively, the *interior*, the *relative interior*, and the *closure* of S, and by X the constraint set for (P), i.e.,

$$X = \{x \in \Re^m \mid Ex \geqq b\}.$$

(For the equality constraint problem (P^{E}), we replace X by $\{x \in \Re^m \mid Ex = b\}$.) We make the following standing assumptions.

* Received by the editors October 7, 1988; accepted for publication (in revised form) May 1, 1989.

† Center for Intelligent Control Systems, Room 35-205, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This research was partially supported by United States Army Research Office contract DAAL03-86-K-0171 (Center for Intelligent Control Systems) and by National Science Foundation grant NSF-ECS-8519058.

Assumption A. f is strictly convex, lower semicontinuous, and continuous within S. Moreover, the *conjugate function* of f defined by

$$f^*(t) = \sup \{\langle t, \xi\rangle - f(\xi) \mid \xi \in \Re^m\} \tag{1.2}$$

is real valued, i.e., $-\infty < f^*(t) < +\infty$ for all $t \in \Re^m$.

Assumption B. $S = S^1 \cap S^2$, where S^1 and S^2 are convex sets in $\Re^m$ such that cl (S^1) is a polyhedral set and $S^1 \cap \text{ri}\,(S^2) \cap X \neq \varnothing$.

Assumption B is the usual feasibility assumption for (P), i.e., $S \cap X \neq \varnothing$, plus an additional constraint qualification. The constraint qualification, which is necessary for establishing the convergence of our algorithm (see § 3), is almost always satisfied. For example, cl (S) (but not necessarily S) is a polyhedral set if f is *separable* [55], [59]. (In this case cl (S) is a box.) Assumption A implies that, for every t, there is some $\xi \in \Re^m$ achieving the supremum in (1.2) and $f(x) \to +\infty$ as $\|x\| \to +\infty$. It follows from the latter that f has bounded level sets. Because f is lower semicontinuous, its level sets are compact. This, together with the feasibility of (P) and the strict convexity of f within S, implies that *there exists a unique optimal solution* to (P), which we denote by x^*.

A dual program of (P), obtained by assigning a Lagrange multiplier p_i to the ith constraint of $Ex \geqq b$, is

$$\text{(D)} \qquad \underset{p \geqq 0}{\text{Maximize}}\ q(p)$$

where

$$\begin{aligned} q(p) &= \min \{f(x) + \langle p, b - Ex\rangle \mid x \in \Re^m\} \\ &= \langle p, b\rangle - f^*(E^T p). \end{aligned} \tag{1.3}$$

(For the equality constraint problem (P^E), the dual program is identical to (D) except that it does not have the nonnegativity constraints on p.) Problem (D) is a concave program with simple nonnegative orthant constraints. Furthermore, strong duality holds for (P) and (D), i.e., the optimal value in (P) equals the optimal value in (D) (see [60, § 1]).

Since f^* is real valued and f is strictly convex, f^* and q are continuously differentiable ([54, Thm. 26.3]). Using the chain rule, we obtain the gradient of q at p, denoted by $d(p)$, to be

$$d(p) = b - E\chi(p), \tag{1.4}$$

where we denote

$$\begin{aligned} \chi(p) &= \nabla f^*(E^T p) \\ &= \arg\max \{\langle p, E\xi\rangle - f(\xi) \mid \xi \in \Re^m\}. \end{aligned} \tag{1.5}$$

We will also denote by $d_i(p)$ the ith coordinate of $d(p)$ and, for any $I \subseteq \{1, \cdots, n\}$, by $d_I(p)$ the vector $(\cdots, d_i(p), \cdots)_{i \in I}$. Note from (1.5) that $\chi(p)$ is also the unique vector x satisfying

$$E^T p \in \partial f(x), \tag{1.6}$$

where $\partial f(x)$ denotes the *subdifferential* of f at x [54].

Note that $p \in \Re^n$ is an optimal solution for (D) if and only if $p = [p + d(p)]^+$, where $[\cdot]^+$ denotes the orthogonal projection onto the nonnegative orthant. However,

(D) is not guaranteed to have an optimal solution. Consider the following example: $n = m = 1$, $E = -1$, $b = 0$, and

$$f(x) = \begin{cases} x^2 - (x)^{1/2} & \text{if } x \geqq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

It can be verified that Assumptions A and B hold, but f does not have a dual support at the optimal primal solution $x^* = 0$.

Differentiability of q motivates a block coordinate ascent method for solving (P) and (D) whereby, given a dual vector p, a block of coordinates are adjusted to increase the dual functional q. Important advantages of such a coordinate ascent method are simplicity, the ability to exploit problem sparsity, and the potential for massive parallelization. As an example, suppose that $f(x)$ is quadratic of the form $\langle x, Qx\rangle/2 + \langle c, x\rangle$, where Q is an $m \times m$ symmetric positive definite matrix and $c \in \Re^m$. Then two coordinates p_i and p_j are uncoupled and can be iterated upon simultaneously if the (i,j)th entry of $EQ^{-1}E^T$ is zero (another example is if f is separable and the (i,j)th entry of EE^T is zero).

Coordinate ascent methods for maximizing general differentiable concave functions have been well studied [5, § 3.2.4], [20], [40], [51], [52], [56], [63], but convergence typically requires compactness of the level sets and some form of strict concavity of the objective function—neither of which holds for q. Coordinate ascent methods for maximizing dual functionals of the form q, on the other hand, have been studied for certain special cases only (e.g., f is differentiable strongly convex, and exact line search is used). More general results are given in [59] and [60], but these results are applicable only for single (not block) coordinate relaxation and for a special type of inexact line search. This lack of a general theory is unfortunate given that dual coordinate ascent methods are among the most popular (and sometimes the only practical) methods for large-scale optimization, e.g., network flow [3], [6], [11], [13], [14], [45], [58], [64], [65], entropy maximization [2], [8], [11], [21], [23], [24], [31]-[33], [37], [41], [48], [57], linear [43] and quadratic programming [12], [15], [18], [27], [28], [38], [39], [42]. Another dual method whose convergence properties are not well understood is the *dual gradient* method [22], [25], [34], [39], [49], [50], [61], [62]. In this method, the dual vector p is moved along the gradient direction (or an approximation of) at each iteration instead of along a coordinate ascent direction. This method can take advantage of second-order derivative information and, in certain cases, is more efficient than dual coordinate ascent methods.

This paper represents an attempt to fill the existing theoretical gaps for the above dual ascent methods. In particular, we (i) propose a general class of (block) coordinate ascent algorithms for maximizing q, (ii) prove various convergence properties for this class, and (iii) show that the dual methods proposed in [2], [6], [8]-[10], [13]-[15], [18], [21], [23], [24], [27], [28], [31]-[33], [37]-[39], [41], [42], [45], [48]-[50], [57]-[60], [62], [64] are in this class. We also present some new algorithms from this class and propose a technique for parallelizing the line search step when f is separable.

This paper is organized as follows: in §§ 2 and 3 we present a coordinate relaxation algorithm and prove that it converges. In § 4 we present an extension of this algorithm for the case where f is strongly convex. In § 5 we consider implementation issues and in § 6 we show that this algorithm contains as special cases a number of known methods. In § 7 we present a technique for parallelizing the line search step in this algorithm when f is separable. In § 8 we give our conclusion and discuss extensions.

We will use the following notation. For any $k \times l$ matrix A, any vector c in $\Re^k$, and any $I \subseteq \{1, \cdots, k\}$, $J \subseteq \{1, \cdots, l\}$, we denote by A_I the matrix $[a_{ij}]_{i \in I, j \in \{1, \cdots, l\}}$, A_{IJ}

the matrix $[a_{ij}]_{i\in I, j\in J}$, and c_I the vector $(\cdots, c_i, \cdots)_{i\in I}$, where a_{ij} is the (i,j)th entry of A and c_i is the ith coordinate of c. For any finite set I and any $J\subseteq I$, we denote by $I\backslash J$ the complement of J relative to I and by $|J|$ the number of elements in J. For any x and z in $\Re^m$, we denote by $f'(x; z)$ the *directional derivative* of f at x along z ([54, pp. 213 and 217]), i.e.,

$$f'(x; z) = \lim_{\lambda\downarrow 0} (f(x+\lambda z) - f(x))/\lambda$$
$$= \max\{\langle z, \eta\rangle \mid \eta \in \partial f(x)\}. \tag{1.7}$$

2. Block coordinate relaxation algorithm. In this section we present our main algorithm, called the *block coordinate relaxation* (BCR) *algorithm*, for solving (P) and (D). In this algorithm, we choose a collection $\mathscr{C}$ of nonempty subsets of $N=\{1, \cdots, n\}$ such that their union equals N and, for each $I\in\mathscr{C}$, we choose continuous functions $\phi_I : \Re^n \times [0, +\infty)^n \to [0, +\infty)$ and $\delta_I : \Re^n \times \Re^n \to [0, +\infty)$ satisfying

(2.1a) $\phi_I(\eta, \pi)$ is bounded away from zero $\Leftrightarrow \pi_I - [\pi_I + \eta_I]^+$ is bounded away from zero,

(2.1b) $\delta_I(\eta, \eta') = 0 \Leftrightarrow \eta_I = \eta'_I$.

We also fix a relaxation parameter $\gamma\in(0, 1]$. Each iteration of the BCR algorithm generates a new dual vector p' from the current dual vector p as follows.

Block Coordinate Relaxation (BCR) *Iteration.*
Given a nonnegative $p\in\Re^n$, choose an $I\in\mathscr{C}$.
Find any nonnegative $p'\in\Re^n$ satisfying

(2.2a) $$p'_i = p_i, \quad \forall i \notin I,$$

(2.2b) $$(\gamma-1)(q(p')-q(p)) \leqq \gamma\langle p'-p, d(p')\rangle,$$

(2.2c) $$\delta_I(d(p'), d(p)) \geqq \phi_I(d(p'), p').$$

Roughly speaking, (2.2a) ensures that only coordinates corresponding to I change value; (2.2b) ensures that a dual ascent occurs; and (2.2c) ensures that the change in the gradient of q, namely $d(p')-d(p)$, is nonzero if p' is not optimal with respect to the coordinates corresponding to I (i.e., $p'_I \neq [p'_I + d_I(p')]^+$). The scalar γ controls the amount of under/over-relaxation in the iteration. ($\gamma>1$ ($\gamma<1$) implies under-relaxation (over-relaxation).)

For solving the equality constraint problem (P^E), we modify the BCR iteration as follows: We replace (2.1a) with "$\phi_I(\eta, \pi)$ is bounded away from zero $\Leftrightarrow$ η is bounded away from zero" and remove the nonnegativity constraints on p and p'. (The extension to mixed equality/inequality constraints is straightforward.)

To ensure that the BCR iteration is well defined (i.e., for any nonnegative $p\in\Re^n$ and $I\in\mathscr{C}$, a p' satisfying (2.2a)-(2.2c) exists), additional assumptions on ϕ_I and δ_I are required. We will see in §§ 5 and 6 that the choice of ϕ_I and δ_I is very important: different choices lead to different methods and, for special cases of (P), the appropriate choice can significantly reduce the work per iteration. We will also see in § 5 that very little needs to be assumed about ϕ_I and δ_I either to make the BCR iteration well defined or to implement it. (For example, if p' is obtained from p by maximizing $q(p)$ over all p_i, $i\in I$, with the other coordinates held fixed, then p' can be shown to satisfy (2.2a)-(2.2c).)

The BCR algorithm that consists of successive applications of the BCR iteration is not guaranteed to converge, unless the coordinates are relaxed in some order. (We

say a coordinate is "relaxed" if the BCR iteration is applied with an $I \in \mathscr{C}$ that contains the index of that coordinate.) We will consider the following two orders of relaxation.

Essentially cyclic order. There exists $T \geqq 1$ such that every coordinate is chosen at least once for relaxation between iterations r and $r+T$, for all $r = 0, 1, \cdots$.

Gauss-Southwell order. At each iteration, choose an $I \in \mathscr{C}$ satisfying

$$\phi_I(d(p), p) \geqq \rho \cdot \max_{J \in \mathscr{C}} \{\phi_J(d(p), p)\},$$

where ρ is a constant in $(0, 1]$.

The above two orders of relaxation are extensions of those discussed in [40, § 7.8] and [56]. We will weaken the essentially cyclic order in § 4. If $\mathscr{C}$ is a partition of N (i.e., the elements of $\mathscr{C}$ are mutually disjoint) and the essentially cyclic order of relaxation is used with $T = |\mathscr{C}| - 1$, then we will say that the order of relaxation is *cyclic.*

3. Main convergence theorem. Let p^r denote the iterate generated by the BCR algorithm at the rth iteration and let $x^r = \chi(p^r)$ $(r = 0, 1, \cdots)$. In this section, we show that, under either the essentially cyclic or the Gauss-Southwell order of relaxation, the BCR algorithm converges, in the sense that $\{x^r\} \to x^*$. We also provide sufficient conditions under which $\{p^r\}$ converges. To simplify the presentation, let I^r denote the set of indexes of the coordinates relaxed at the rth iteration and let $d^r = d(p^r)$ $(r = 0, 1, \cdots)$. Our argument will follow closely that in § 3 of [60] (in fact, to simplify the presentation, we will borrow certain results from [60]).

We precede our proof of convergence with the following four technical lemmas.

LEMMA 1. (a) *For any y in S and any sequence of vectors $\{y^1, y^2, \cdots\}$ in S such that $\{f(y^k) + f'(y^k; y - y^k)\}$ is bounded from below, it holds that both $\{y^k\}$ and $\{f(y^k)\}$ are bounded, and every limit point of $\{y^k\}$ is in S.*

(b) *For any $y \in S$, any $z \in \Re^m$ such that $y + z \in S$, and any sequences $\{y^1, y^2, \cdots\} \to y$ and $\{z^1, z^2, \cdots\} \to z$ such that $y^k \in S$ and $y^k + z^k \in S$ for all k, it holds that*

$$\lim_{k \to +\infty} \sup \{f'(y^k; z^k)\} \leqq f'(y; z).$$

(c) *For any $y \in S$, there exists a positive scalar ε such that $\{x \in S \mid \|x - y\| \leqq \varepsilon\}$ is closed.*

Proof. Parts (a) and (b) follow from, respectively, the proofs of Lemmas 2 and 3 in [60]. Part (c) follows from the proof of Proposition 1(b) in [60]. □

LEMMA 2. *For $r = 0, 1, \cdots$,*

$$q(p^{r+1}) - q(p^r) \geqq \gamma[f(x^{r+1}) - f(x^r) - f'(x^r; x^{r+1} - x^r)], \tag{3.1a}$$

$$f(x^r) + f'(x^r; x^* - x^r) \geqq q(p^r). \tag{3.1b}$$

Proof. We first prove (3.1). From (1.2), (1.3), and (1.5) we have that

$$q(p^r) = \langle p^r, b\rangle + f(x^r) - \langle E^T p^r, x^r\rangle, \quad \forall r = 0, 1, \cdots. \tag{3.2a}$$

Since (cf. (1.4)) $d^r = b - Ex^r$, this implies that, for any r,

$$q(p^{r+1}) - q(p^r) = f(x^{r+1}) - f(x^r) - \langle E^T p^r, x^{r+1} - x^r\rangle + \langle p^{r+1} - p^r, d^{r+1}\rangle. \tag{3.2b}$$

Multiplying both sides by $(\gamma - 1)$ and using (2.2b), we obtain

$$\begin{aligned}\gamma\langle p^{r+1} - p^r, d^{r+1}\rangle &\geqq (\gamma - 1)[q(p^{r+1}) - q(p^r)] \\ &= (\gamma - 1)[f(x^{r+1}) - f(x^r) - \langle E^T p^r, x^{r+1} - x^r\rangle] \\ &\quad + (\gamma - 1)\langle p^{r+1} - p^r, d^{r+1}\rangle.\end{aligned}$$

Hence

$$\langle p^{r+1}-p^r, d^{r+1}\rangle \geqq (\gamma-1)[f(x^{r+1})-f(x^r)-\langle E^T p^r, x^{r+1}-x^r\rangle],$$

which, together with (3.2b), implies that

$$\begin{aligned} q(p^{r+1})-q(p^r) &\geqq \gamma[f(x^{r+1})-f(x^r)-\langle E^T p^r, x^{r+1}-x^r\rangle] \\ &\geqq \gamma[f(x^{r+1})-f(x^r)-f'(x^r; x^{r+1}-x^r)], \end{aligned}$$

where the second inequality follows from (1.7) and the fact (cf. (1.6)) $E^T p^r \in \partial f(x^r)$.

To prove (3.1b), note that since $p^r \geqq 0$ and x^* satisfies (1.1), then

$$\begin{aligned} q(p^r) &\leqq q(p^r)+\langle p^r, Ex^*-b\rangle \\ &= f(x^r)+\langle E^T p^r, x^*-x^r\rangle \\ &\leqq f(x^r)+f'(x^r; x^*-x^r), \end{aligned}$$

where the equality follows from (3.2a) and the second inequality follows from (1.7) and the fact (cf. (1.6)) $E^T p^r \in \partial f(x^r)$. □

Note that, by (3.1a), the sequence $\{q(p^r)\}$ is monotonically increasing. Lemma 1(a), (b) and Lemma 2 imply the following two lemmas.

LEMMA 3. (a) $\{x^r\}$ *and* $\{f(x^r)\}$ *are bounded and each limit point of* $\{x^r\}$ *is in* S.

(b) $\{x^{r+1}-x^r\} \to 0$.

(c) $\{p^r_{I^r}-[p^r_{I^r}+d^r_{I^r}]^+\} \to 0$.

Proof. Since $\{q(p^r)\}$ is monotonically increasing and x^r is in S for all r, (3.1b) and Lemma 1(a) imply part (a). Now, if part (b) does not hold, then there must exist a subsequence R for which $\{x^r\}_{r\in R}$ converges to some point x^∞ and $\{x^{r+1}\}_{r\in R}$ converges to some point $x'' \neq x^\infty$. Let $z = x''-x^\infty$ $(z \neq 0)$. By part (a), both x^∞ and $x^\infty+z$ are in S. Then from (3.1a), the continuity of f on S, and Lemma 1(b), we obtain

$$\lim_{r\to+\infty, r\in R} \inf\{q(p^{r+1})-q(p^r)\} \geqq \gamma[f(x^\infty+z)-f(x^\infty)-f'(x^\infty; z)].$$

Since $q(p^r)$ is nondecreasing with r and f is strictly convex (so the right-hand side of above is a *positive* scalar), it follows that

$$\{q(p^r)\} \to +\infty.$$

This, in view of the strong duality condition

$$\max\{q(p) \mid p \geqq 0\} = \min\{f(x) \mid Ex \geqq b\},$$

contradicts the feasibility of (P), i.e., $S \cap X \neq \emptyset$.

If part (c) does not hold, then there exist scalar $\varepsilon > 0$, coordinate block $I \in \mathscr{C}$, and subsequence R for which

$$I^r = I \quad \text{and} \quad \|p^r_I-[p^r_I+d^r_I]^+\| \geqq \varepsilon, \quad \forall r \in R.$$

Then (2.1a) implies that $\{\phi_I(d^r, p^r)\}_{r\in R}$ is bounded away from zero, i.e., there exists some scalar $\theta > 0$ such that

$$\phi_I(d^r, p^r) \geqq \theta, \quad \forall r \in R.$$

It follows from (2.2c) that

$$\delta_I(d^r, d^{r-1}) \geqq \theta, \quad \forall r \in R. \tag{3.3}$$

Since (cf. (1.4)) $d^r = b - Ex^r$, $\{d^r\}$ is bounded by part (a). This, together with (2.1b), (3.3), and the continuity of δ_I, implies that

$$\|E_I(x^r - x^{r-1})\| = \|d_I^r - d_I^{r-1}\| \geqq \theta', \quad \forall r \in R,$$

for some scalar $\theta' > 0$. This contradicts part (b). □

LEMMA 4. *Under either the essentially cyclic or the Gauss-Southwell order of relaxation, if x^∞ is any limit point of $\{x^r\}$, then $x^\infty \in S \cap X$ and there exists a subsequence $\{x^r\}_{r\in R} \to x^\infty$ satisfying*

$$b_i - E_i x^\infty < 0 \Rightarrow \{p_i^r\}_{r\in R} \to 0. \tag{3.4}$$

Proof. We will first prove that

$$\{p^r - [p^r + d^r]^+\} \to 0. \tag{3.5}$$

Suppose that the essentially cyclic order is used. Fix any coordinate index $i \in N$ and, for each $r \geqq T$, let $\tau(r)$ be the largest integer h not exceeding r such that $i \in I^{h-1}$. Then

$$d_i^r = d_i^{\tau(r)} + \sum_{h=\tau(r)}^{r-1} E_i(x^h - x^{h+1}), \quad \forall r \geqq T.$$

Since $r - \tau(r) \leqq T$ for all $r \geqq T$, this, together with Lemma 3(b), (c) and the fact $p_i^r = p_i^{\tau(r)}$ for all $r \geqq T$, implies (3.5). Now suppose that the Gauss-Southwell order is used. Then (2.1a) and Lemma 3(c) imply that $\{p_I^r - [p_I^r + d_I^r]^+\} \to 0$ for all $I \in \mathscr{C}$. Since the union of the elements of $\mathscr{C}$ equals N, (3.5) holds.

Since $|p_i^r - [p_i^r + d_i^r]^+| = d_i^r$ if $d_i^r \geqq 0$, it follows from (3.5) that $\lim_{r\to+\infty} \sup\{d_i^r\} \leqq 0$ for all i. Hence every limit point of $\{x^r\}$ is in X. This, together with Lemma 3(a), implies that $x^\infty \in S \cap X$.

Next we prove (3.4). Let $d = b - Ex^\infty$. Since $x^\infty \in X$, we have $d_i \leqq 0$ for all i. Consider any i such that $d_i < 0$ (if no such i exists, we are done). Since x^∞ is a limit point of $\{x^r\}$, there exists subsequence R such that $\{x^r\}_{r\in R} \to x^\infty$. Then $\{d_i^r\}_{r\in R} \to d_i < 0$, which, together with (cf. (3.5)) $\{p_i^r - [p_i^r + d_i^r]^+\}_{r\in R} \to 0$, implies that $\{p_i^r\}_{r\in R} \to 0$. □

Lemmas 1 and 4 allow us to prove the main result of this section.

PROPOSITION 1. *If $\{p^r\}$ is a sequence of dual vectors generated by the* BCR *algorithm under either the essentially cyclic or the Gauss-Southwell order of relaxation, then the following hold*:

(a) $\{\chi(p^r)\} \to x^*$.

(b) *If* cl (S) *is a polyhedral set, and there exists a closed ball B around x^* such that $f'(x; (y-x)/\|y-x\|)$ is bounded for all x, y in $B \cap S$, then $\{q(p^r)\} \to f(x^*)$.*

(c) If int $(X) \cap S \neq \varnothing$, *then $\{p^r\}$ is bounded and every one of its limit points is an optimal solution for* (D).

Proof. We prove part (a) only. The proof of parts (b) and (c) is identical to that of Proposition 1 in [60]. Let $x^r = \chi(p^r)$ for all r, let x^∞ be a limit point of $\{x^r\}$, and let R be a subsequence of $\{1, 2, \cdots\}$ satisfying (3.4). Also let $d = b - Ex^\infty$ and $I^- = \{i \in N \mid d_i < 0\}$. By Lemma 4, $x^\infty \in S \cap X$. Suppose that $x^\infty \neq x^*$ and we will reach a contradiction.

Let y be any element of $S^1 \cap \text{ri}\,(S^2) \cap X$ (y exists by Assumption B). Fix any $\lambda \in (0, 1)$ and denote $y(\lambda) = \lambda y + (1-\lambda)x^*$. Then $y(\lambda) \in S^1 \cap \text{ri}\,(S^2) \cap X$. By Lemma 1(c), there exists an $\varepsilon > 0$ such that $\{x \in S \mid \|x - x^\infty\| \leqq \varepsilon\}$ is closed. Since cl (S^1) is a polyhedral set and $y(\lambda) - x^\infty$ belongs to the tangent cone of S^1 at x^∞, this implies that there exists $\delta \in (0, 1)$ such that, for all $r \in R$ sufficiently large,

$$x^r + \delta z \in S^1, \tag{3.6}$$

where $z = y(\lambda) - x^\infty$. On the other hand, since $y(\lambda) \in \text{ri}\,(S^2)$, $x^r \in S^2$ for all r, and $\{x^r\}_{r\in R} \to x^\infty$, we have that, for all $r \in R$ sufficiently large,

$$x^r + \delta z \in S^2. \tag{3.7}$$

Since $y(\lambda) \in X$, $E_i z \geqq 0$ for all $i \notin I^-$. This implies that (since $p^r \geqq 0$)

$$\langle p^r, Ez\rangle \geqq \sum_{i\in I^-} p_i^r (E_i z), \quad \forall r \in R, \quad \text{if } I^- \neq \varnothing,$$

$$\langle p^r, Ez\rangle \geqq 0, \quad \forall r \in R, \quad \text{otherwise.}$$

In either case, we have (cf. (3.4))

$$\lim_{r\to+\infty, r\in R} \inf\{\langle p^r, Ez\rangle\} \geqq 0. \tag{3.8}$$

Since $x^\infty + \delta z \in S$ and (cf. (1.7) and the fact $E^T p^r \in \partial f(x^r))$ $f'(x^r; z) \geqq \langle p^r, Ez\rangle$ for all r, (3.6)-(3.8) and Lemma 1(b) imply that

$$f'(x^\infty; z) \geqq 0.$$

Hence $f(x^\infty) \leqq f(y(\lambda))$. Since the choice of $\lambda \in (0, 1)$ was arbitrary, by taking λ arbitrarily small (and using the continuity of f within S), we obtain that $f(x^\infty) \leqq f(x^*)$. Since f is strictly convex and $x^\infty \in S \cap X$, this contradicts the hypothesis $x^\infty \neq x^*$. □

We remark that a result analogous to Proposition 1 also holds for the BCR algorithm modified to solve the equality constraint problem (P^E).

Extensions.

1. Note from its proof that Proposition 1 still holds if Assumption B is replaced by the following more general assumption: $S \cap X \neq \varnothing$ and, for any $x \in S \cap X$, any $y \in S \cap X$, and any sequence $\{y^k\}$ in S such that $\{y^k\} \to y$, it holds $f'(y; x - y) \geqq \lim_{k\to+\infty} \sup\{f'(y^k; x - y)\}$.

2. Consider a *mixed* algorithm whereby in between BCR iterations are inserted other dual ascent iterations (these other dual ascent iterations need not be convergent on their own). For this mixed algorithm, Proposition 1 can also be shown to hold, provided that (i) there exists a bound $T' \geqq 1$ such that a BCR iteration is executed at least once every T' consecutive iterations (the coordinates are assumed to be relaxed by the BCR iterations in either the essentially cyclic or the Gauss-Southwell order) and that (ii) each inserted dual ascent iteration generates the new dual vector p' from the current dual vector p with the property

$$q(p') - q(p) \geqq \mu[f(\chi(p')) - f(\chi(p)) - f'(\chi(p); \chi(p') - \chi(p))], \tag{3.9}$$

where μ is some fixed positive scalar. (The condition (3.9) is satisfied by most dual ascent iterations. For example, it is satisfied by the BCR iteration (cf. (3.1b)) as well as by any iteration that maximizes exactly q along some direction in $\Re^n$ (cf. (3.2b)).) In such a mixed algorithm, the BCR iteration may be viewed as a *spacer* step to enforce convergence of the iterates. We give an application of this mixed algorithm below. Other applications are discussed in § 6.2. Other extensions of the BCR algorithm are discussed in Proposition 10 and in § 8.

3. It is easily seen that every limit point of the sequence $\{p^r\}$ is an optimal solution of (D). Hence $\{p^r\}$ diverges if (D) does not have an optimal solution. On the other hand if the set of optimal solutions for (D) is nonempty but unbounded, $\{p^r\}$ can still diverge (and thus cause numerical difficulty). To remedy this, we can modify the BCR algorithm by, for example, replacing p with

$$p' = \text{argmin}\{\|\pi\| \mid E^T\pi = E^T p, \langle b, \pi\rangle = \langle b, p\rangle, \pi \geqq 0\} \tag{3.10}$$

at periodic intervals. Because $E^T p' = E^T p$ and $\langle b, p' \rangle = \langle b, p \rangle$, it follows from (1.3) and (1.5) that $q(p') = q(p)$ and $\chi(p') = \chi(p)$. Hence iteration (3.10) satisfies (3.9) and the modified BCR algorithm is convergent. In some cases (e.g., network flow), the minimization (3.10) can be performed quite efficiently.

4. Convergence for strongly convex costs. In this section we consider the special case where f is *strongly convex*, in the sense that there exist scalars $\sigma > 0$ and $\omega > 1$ such that

$$f(y) - f(x) - f'(x; y-x) \geqq \sigma \|y-x\|^{\omega}, \quad \forall x \in S, \quad \forall y \in S. \tag{4.1}$$

We show that, for this special case, the essentially cyclic order of relaxation can be weakened further. (Note that (4.1) is a generalization of the traditional definition of strong convexity (called *uniform convexity* in [47, p. 83]), where ω is taken to be 2. As an example, $f: \Re \to (-\infty, +\infty]$ given by

$$f(x) = \begin{cases} x^4 & \text{if } x \geqq 0, \\ +\infty & \text{otherwise,} \end{cases}$$

satisfies (4.1) with $\omega = 4$, $\sigma = \frac{1}{4}$, but does not satisfy (4.1) with $\omega = 2$ for any positive σ.)

Consider the following order of relaxation that is weaker than the essentially cyclic order.

Quasi-cyclic order. Every coordinate is chosen at least once for relaxation between iterations τ_k and τ_{k+1}, for $k = 1, 2, \cdots$, where $\{\tau_1, \tau_2, \cdots\}$ is a sequence of integers given by

$$\tau_1 = 0 \quad \text{and} \quad \tau_{k+1} = \tau_k + b_k, \qquad k = 1, 2, \cdots,$$

and $\{b_k\}$ is any sequence of integers satisfying

$$b_k \geqq |\mathscr{C}|, \qquad k = 1, 2, \cdots, \quad \text{and} \quad \sum_{k=1}^{\infty} (b_k)^{1-\omega} = +\infty.$$

(For example, $b_k = n \cdot k^{1/(\omega-1)}$ is a valid choice.) The above order of relaxation, first proposed in [59], is similar to the essentially cyclic order but allows the length of the cycle to grow without bound. It is an open question whether this order can be weakened further, say, to one that only assumes that each coordinate is relaxed an infinite number of times.

By using Lemmas 1 and 3 in § 3 and an argument analogous to that for Proposition 2 in [60], we can show the following result that is analogous to Proposition 1 (which, for simplicity, we state without proof).

PROPOSITION 2. *Suppose that* (4.1) *holds and let* $\{p^r\}$ *be a sequence of dual vectors generated by the* BCR *algorithm under the quasi-cyclic order of relaxation. Then the following hold*:

(a) *There exists a subsequence* $R \subseteq \{1, 2, \cdots\}$ *such that* $\{\chi(p^r)\}_{r \in R} \to x^*$.

(b) *If* cl (S) *is a polyhedral set, and there exists a closed ball* B *around* x^* *such that* $f'(x; (y-x)/\|y-x\|)$ *is bounded for all* x, y *in* $B \cap S$, *then* $\{q(p^r)\} \to f(x^*)$ *and* $\{\chi(p^r)\} \to x^*$.

(c) *If* int $(X) \cap S \neq \varnothing$, *then* $\{q(p^r)\} \to f(x^*)$, $\{\chi(p^r)\} \to x^*$, *and* $\{p^r\}$ *is bounded. Moreover, each limit point of* $\{p^r\}$ *is an optimal solution for* (D).

Note that the conclusion of Proposition 2(a) is weaker than that of Proposition 1(a). Only for the special case where f is separable and $\mathscr{C} = \{\{1\}, \cdots, \{n\}\}$ has it been shown that $\{x^r\} \to x^*$, assuming only that (4.1) holds and that the quasi-cyclic order of relaxation is used [59].

5. Choosing ϕ_I and δ_I. We have seen from §§ 3 and 4 that the BCR algorithm converges, provided that each BCR iteration is well defined. In this section we will consider choices of ϕ_I and δ_I that ensure that the BCR iteration is well defined. In particular, we will show that it is well defined if either a certain constraint qualification holds or if ϕ_I and δ_I satisfy a certain growth condition. We will also consider particular implementations of the BCR iteration for, respectively, single coordinate relaxation and gradient ascent.

5.1. Exact coordinate maximization. In this subsection we show that the BCR iteration (2.2a)-(2.2c) is well defined if the dual functional $q(p)$ can be maximized *exactly* with respect to the coordinates p_i, $i \in I$, while the other coordinates are held fixed. To ensure that such an exact maximization is possible, the following constraint qualification on f and X will be considered.

Assumption C. The function f is of the form $f = h + \delta_C$, where $h : \Re^m \to (-\infty, +\infty]$ is a strictly convex function such that $\text{ri}\,(\text{dom}\,(h)) \cap X \neq \varnothing$ and δ_C is the indicator function for a polyhedral set C in $\Re^m$ (i.e., $\delta_C(x) = 0$ if $x \in C$ and $\delta_C(x) = +\infty$ otherwise).

We have the following result.

PROPOSITION 3. *Under Assumption* C (*in addition to Assumptions* A *and* B), *the* BCR *iteration* (2.2a)-(2.2c) *with* $\gamma \in (0, 1]$ *is well defined for any functions* ϕ_I *and* δ_I *satisfying, respectively,* (2.1a) *and* (2.1b).

Proof. Consider any nonnegative vector $p \in \Re^n$ and any $I \in \mathscr{C}$. Let us define the following relaxed problem

$$\begin{aligned} &\text{Minimize} \quad f(x) - \sum_{i \notin I} p_i(E_i x) \\ &\text{subject to} \quad E_I x \geqq b_I. \end{aligned} \tag{5.1}$$

Since (P) is feasible, so is (5.1). Since the cost function of (5.1) has compact level sets (cf. Assumption A), this implies that (5.1) has an optimal solution which we denote by x'. It then follows from Assumption C and Theorem 28.2 in [54] that x', together with some Lagrange multiplier vector associated with the constraints $E_I x \geqq b_I$, satisfies the Kuhn-Tucker conditions for (5.1). Let Δ_i, $i \in I$, denote the coordinate of this Lagrange multiplier vector associated with the constraint $E_i x \geqq b_i$. Define a new dual vector $p' \in \Re^n$ to be the vector whose ith coordinate is

$$p'_i = \begin{cases} \Delta_i & \text{if } i \in I, \\ p_i & \text{otherwise.} \end{cases} \tag{5.2}$$

We claim that $p' \geqq 0$ and satisfies (2.2b) and (2.2c) (clearly p' satisfies (2.2a)). To see this, note from the Kuhn-Tucker conditions for (5.1) that $\Delta_i \geqq 0$, for all $i \in I$, and

$$E_i x' = b_i \quad \text{if } \Delta_i > 0, \qquad i \in I, \tag{5.3a}$$

$$E_i x' \geqq b_i \quad \text{if } \Delta_i = 0, \qquad i \in I, \tag{5.3b}$$

$$\sum_{i \in I} (E_i)^T \Delta_i \in \partial f(x') - \sum_{i \notin I} (E_i)^T p_i. \tag{5.3c}$$

Hence $p' \geqq 0$ and (by (1.6), (5.2), (5.3c)) $\chi(p') = x'$. This, together with (1.4) and (5.2), implies that

$$\begin{aligned} \langle p' - p, d(p') \rangle &= \langle p' - p, b - Ex' \rangle \\ &= \sum_{i \in I} (\Delta_i - p_i)(b_i - E_i x') \\ &\geqq 0, \end{aligned}$$

where the inequality follows from (5.3a), (5.3b). Therefore, $q(p') \geqq q(p)$. Since $\gamma \in (0, 1]$, this implies that (2.2b) holds.

To see that (2.2c) holds, note that (cf. (5.3a), (5.3b) and the fact $d_I(p') = b_I - E_I x'$) $p'_I = [p'_I + d_I(p')]^+$. Hence, by (2.1a), $\phi_I(d(p'), p') = 0$. □

From the proof of Proposition 3 it can be seen that the dual vector p' given by (5.2) is obtained equivalently by maximizing exactly the dual functional $q(p)$ with respect to p_I, while the other coordinates of p are held fixed. If $q(p)$ has bounded level sets and is strictly convex in p_I for all I in $\mathscr{C}$, then the convergence of such a *nonlinear Gauss-Seidel* iteration follows from Proposition 2.5 in [5, § 3.2.4]. However, the conditions under which q has this property are very restrictive.

It is an open question whether the hypothesis in Proposition 3 can be weakened further. For example, would the conclusion of Proposition 3 hold if it is only assumed that (D) has an optimal solution?

5.2. ϕ_I and δ_I satisfy a growth condition. If Assumption C does not hold, then we need to impose some growth conditions on ϕ_I and δ_I to ensure that the BCR iteration is well defined. We state this result in the following proposition.

PROPOSITION 4. *If in the* BCR *iteration, it holds that*

$$\phi_I(\eta, \pi) \leqq \|\pi_I - [\pi_I + \eta_I]^+\|, \quad \forall \eta \in \Re^n, \quad \forall \pi \in [0, +\infty)^n,$$

$$\delta_I(\eta, \eta') \geqq \|\eta_I - \eta'_I\|, \quad \forall \eta, \eta' \in \Re^n,$$

then the BCR *iteration with* $\gamma \in (0, 1]$ *is well defined.*

Proof. Let $p \in \Re^n$ and $I \in \mathscr{C}$ be as in the BCR iteration and let $\beta = \phi_I(d(p), p)$. If $\beta = 0$, then the BCR iteration is well defined (since $p' = p$ satisfies (2.2a)-(2.2c)). Suppose that $\beta > 0$. Let $\theta_i = d_i(p)$ and $I^- = \{i \in I \mid \theta_i < 0\}$, $I^+ = \{i \in I \mid \theta_i > 0\}$. Let μ be any scalar in $(0, \frac{1}{2}]$.

Consider the following relaxed problem

$$\begin{aligned} &\text{maximize} && f(x) - \sum_{i \notin I^-} p_i(E_i x) \\ &\text{subject to} && E_i x \geqq b_i - \theta_i \mu, \quad \forall i \in I^+, \\ & && E_i x \geqq b_i, \quad \forall i \in I^-. \end{aligned} \tag{5.4}$$

First note that the interior of the constraint set for (5.4) intersects S. To see this, let $\zeta(\lambda) = \lambda\chi(p) + (1-\lambda)x^*$. Then (since $\theta_i = b_i - E_i\chi(p)$)

$$E_i\zeta(\lambda) - b_i = \lambda(-\theta_i) + (1-\lambda)(E_i x^* - b_i) \geqq -\lambda\theta_i, \quad \forall i \in I^+,$$

$$E_i\zeta(\lambda) - b_i = \lambda(-\theta_i) + (1-\lambda)(E_i x^* - b_i) \geqq -\lambda\theta_i > 0, \quad \forall i \in I^-,$$

so that, for λ sufficiently small, $\zeta(\lambda)$ is in the interior of the feasible set for (5.4). On the other hand, since $\chi(p)$ and x^* are both in S and S is convex, $\zeta(\lambda) \in S$ for all $\lambda \in [0, 1]$.

Since the interior of the constraint set for (5.4) intersects S, the convex program (5.4) is strictly consistent ([54, p. 300]). It then follows from Assumption A and Corollary 29.1.5 of [54] that there exist an $x' \in \Re^m$ and a Lagrange multiplier vector associated with the constraints in (5.4) that, together, satisfy the Kuhn-Tucker conditions for (5.4). Let Δ_i, $i \in I^+$ ($i \in I^-$), denote the coordinate of this Lagrange multiplier vector associated with the constraint $E_i x \geqq b_i - \theta_i\mu$ ($E_i x \geqq b_i$). Define $p' \in \Re^n$ to be the

vector whose ith coordinate is

$$p_i' = \begin{cases} \Delta_i & \text{if } i \in I^-, \\ p_i + \Delta_i & \text{if } i \in I^+, \\ p_i & \text{otherwise.} \end{cases} \tag{5.5}$$

We claim that $p' \geqq 0$ and satisfies (2.2b) and (2.2c) (clearly p' satisfies (2.2a)). To see this, note from the Kuhn-Tucker conditions for (5.4) that $\Delta_i \geqq 0$, for all $i \in I^+ \cup I^-$, and

$$E_i x' = b_i - \theta_i \mu \quad \text{if } \Delta_i > 0, \qquad i \in I^+, \tag{5.6a}$$

$$E_i x' = b_i \quad \text{if } \Delta_i > 0, \qquad i \in I^-, \tag{5.6a}$$

$$E_i x' \geqq b_i \quad \text{if } \Delta_i = 0, \qquad i \in I^-, \tag{5.6b}$$

$$\sum_{i \in I^+ \cup I^-} (E_i)^T \Delta_i \in \partial f(x') - \sum_{i \notin I^-} (E_i)^T p_i. \tag{5.6c}$$

Hence $p' \geqq 0$ and (by (1.6), (5.5), (5.6c)) $\chi(p') = x'$. The latter implies that (cf. (1.4)) $d_i(p') = b_i - E_i x'$ for all i. This, together with (5.5), (5.6a), (5.6b), and the positivity of $\theta_i \mu$, implies that

$$\begin{aligned} \langle p' - p, d(p') \rangle &= \langle p' - p, b - Ex' \rangle \\ &= \sum_{i \in I^-, \Delta_i = 0} (\Delta_i - p_i)(b_i - E_i x') + \sum_{i \in I^+, \Delta_i > 0} \Delta_i \theta_i \mu \\ &\geqq 0. \end{aligned}$$

Hence $q(p') \geqq q(p)$. Since $\gamma \in (0, 1]$, this implies that (2.2b) holds.

We now show that (2.2c) holds. First note from (5.6a), (5.6b) that $\Delta_i = [\Delta_i + b_i - E_i x']^+$ for all $i \in I^-$. Hence (cf. (5.5))

$$p_i' - [p_i' + d_i(p')]^+ = 0, \quad \forall i \in I^-. \tag{5.7}$$

From the nonexpansive property of $[\cdot]^+$ and $p' \geqq 0$, we also have

$$|p_i' - [p_i' + d_i(p')]^+| \leqq |d_i(p')|, \quad \forall i \in I. \tag{5.8}$$

Since $0 = d_i(p)$ for all $i \in I \setminus (I^- \cup I^+)$, (5.8) implies

$$|p_i' - [p_i' + d_i(p')]^+| \leqq |d_i(p) - d_i(p')|, \quad \forall i \in I \setminus (I^- \cup I^+). \tag{5.9}$$

For each $i \in I^+$, since $d_i(p') \leqq \theta_i \mu$ and $\mu \in (0, \frac{1}{2}]$, we have from $\theta_i = d_i(p)$ that

$$|d_i(p')| \leqq |d_i(p) - d_i(p')|, \quad \forall i \in I^+,$$

and hence (cf. (5.8))

$$|p_i' - [p_i' + d_i(p')]^+| \leqq |d_i(p) - d_i(p')|, \quad \forall i \in I^+. \tag{5.10}$$

Combining (5.7), (5.9), and (5.10), we obtain that $\|p_I' - [p_I' + d_I(p')]^+\| \leqq \|d_I(p) - d_I(p')\|$, which, together with our hypothesis, implies (2.2c). □

The proof of Proposition 4 also suggests an implementation of the BCR iteration—by way of solving (5.4). In fact, from the proof of Proposition 4 we see that (5.4) can be solved *inexactly*, i.e., it suffices to find any nonnegative $p' \in \Re^n$ for which

$$\begin{aligned} &d_i(p') \leqq 0, \quad \forall i \in I^-, \\ &d_i(p') \leqq \delta d_i(p), \quad \forall i \in I^+, \\ &p_i' = p_i, \quad \forall i \notin (I^+ \cup I^-), \end{aligned}$$

where δ is any fixed scalar in $(0, 1)$ (this corresponds to choosing $\phi_I(\eta, \pi) = \|\pi_I - [\pi_I + \eta_I]^+\|$ and $\delta_I(\eta, \eta') = \max\{1, \delta/(1-\delta)\} \cdot \|\eta_I - \eta'_I\|$). With this implementation, the BCR algorithm can be thought of as solving (inexactly) a sequence of subproblems of the form (5.4). The fact that (5.4) can be solved inexactly makes this implementation quite practical.

5.3. Single coordinate relaxation. By choosing the coordinate blocks so that any two coordinates from different blocks are weakly coupled, the BCR algorithm can perform substantially faster than its single coordinate counterpart (the amount of improvement depends on the computational effort per iteration). Nevertheless, for problems that are large and sparse, single coordinate algorithms are often favoured; they are simpler to implement, use less storage, can readily exploit problem sparsity, and converge quite fast. In fact, most of the dual coordinate ascent algorithms are single coordinate algorithms (see § 6).

We will presently consider a specialization of the BCR iteration for the single coordinate case, i.e., $\mathscr{C} = \{\{1\}, \cdots, \{n\}\}$, that is both simple and powerful. For each $i \in N$, let α_i be any scalar in $(0, 1)$ and let $\psi_i : \Re \to \Re$ be any continuous, *strictly increasing* function satisfying $\psi_i(0) = 0$. Consider the following iteration that generates a new dual vector p' from the current dual vector p (e^s denotes the sth coordinate vector in $\Re^n$).

Single Coordinate Relaxation (SCR) *Iteration.*

Given a nonnegative $p \in \Re^n$, choose any $s \in N$ and let $\beta = \psi_s(d_s(p))$.

Set $p' = p + \lambda e^s$, where λ is any scalar satisfying

(5.11a) $\quad \alpha_s\beta \geqq \psi_s(d_s(p')) \geqq 0 \quad$ if $\beta \geqq 0$,

(5.11b) $\quad \alpha_s\beta \leqq \psi_s(d_s(p')) \leqq 0 \quad$ if $\beta < 0 \quad$ and $\quad \psi_s(d_s(p - p_s e^s)) \leqq \alpha_s\beta$,

(5.11c) $\quad \lambda = -p_s \quad$ otherwise.

(For the equality constraint problem (P^E), we modify the SCR iteration as follows: We replace (5.11b), (5.11c) by "$\alpha_s\beta \leqq \psi_s(d_s(p')) \leqq 0$ if $\beta < 0$" and remove the nonnegativity constraint on p.) To see that the stepsize λ is well defined, note that $d_s(p)$ is nondecreasing in p_s (since q is concave and, by (1.4), $d_s(p) = \partial q(p)/\partial p_s$) and ψ_s is strictly increasing. Hence $\lambda > 0$ (< 0) if $\beta > 0$ ($\beta < 0$) and is well defined when it is given by either (5.11b) or (5.11c). If λ is not well defined when it is given by (5.11a), it must be that $\psi_s(d_s(p + \theta e^s)) > \alpha_s\beta$ for all $\theta \geqq 0$. This, together with the properties of ψ_s, implies that $d_s(p + \theta e^s) \geqq \varepsilon$ for all $\theta \geqq 0$, where ε is some positive scalar. Hence

$$\lim_{\theta \to +\infty} q(p + \theta e^s) = +\infty,$$

a contradiction of the feasibility of (P).

Now we show that the SCR iteration is a special case of the BCR iteration with $I = \{s\}$, $\gamma = 1$, $\phi_s(\eta, \pi) = (1/\alpha_s - 1)|[\pi_s + \psi_s(\eta_s)]^+ - \pi_s|$, and $\delta_s(\eta, \eta') = |\psi_s(\eta_s) - \psi_s(\eta'_s)|$. Since $\lambda > 0$ (< 0) if $\beta > 0$ ($\beta < 0$), it follows from (5.11a)-(5.11c) and the properties of ψ_s that $\lambda d_s(p') \geqq 0$. Since $\langle d(p'), p' - p\rangle = \lambda d_s(p')$, this implies that p' satisfies (2.2b) with $\gamma = 1$. Also from (5.11a)-(5.11c) we have that

either $\quad \alpha_s|\beta - \psi_s(d_s(p'))| \geqq (1 - \alpha_s)|\psi_s(d_s(p'))|$

or $\quad p'_s = 0, \quad \psi_s(d_s(p')) < 0,$

which, together with the nonexpansive property of $[\cdot]^+$, implies that

either $\quad \delta_s(d(p'), d(p)) \geqq \phi_s(d(p'), p')$

or $\quad \phi_s(d(p'), p') = 0.$

Hence (2.2c) holds.

Since the SCR iteration is a special case of the BCR iteration, it follows that the algorithm based on successive applications of the SCR iteration converges (in the sense of either Proposition 1 or Proposition 2).

Notes and extensions.

1. If Assumption C holds, then $\alpha_s = 0$ is also allowable (in this case the choice of ψ_s is inconsequential). This is because the SCR iteration with $\alpha_s = 0$ is equivalent to (5.2) with $I = \{s\}$. In this case we obtain that $p'_s = [p'_s + d_s(p')]^+$ and the SCR iteration can be interpreted as an *exact* line search along the sth coordinate direction. We will see in § 6 that most of the single coordinate relaxation methods use exact line search (see [2], [8]-[10], [13], [21], [24], [27], [28], [31]-[33], [39], [41], [45], [48], [50], [57], [58], [64]).

2. In the SCR iteration, λ is always between 0 and the line search stepsize; hence the SCR iteration uses *under-relaxation.* It is possible to also use over-relaxation (i.e., λ exceeding the line search stepsize), if a condition analogous to (2.2b) is imposed.

3. General techniques for computing the stepsize λ in the SCR iteration can be found in [7], [35], [44], [59] (see also [26], [46] for the special case where f is separable and quadratic over a box). In some very special cases, λ can be computed very easily (see § 6.3). If f is separable, then λ can be computed in parallel (see § 7).

5.4. Dual gradient iteration. Consider the equality constrained problem (P^{E}). A classical method for solving this problem is the dual gradient method, whereby at each iteration, the dual vector p is moved along the gradient direction $\nabla q(p)$ (or an approximation of) in order to maximize the dual functional q. This method was one of the first dual methods proposed to solve (P^{E}) [22], [61] (also see [4], [25], [34], [39], [40], [49], [50]) and yet, despite its long history, very little is known about its convergence properties. In this subsection, we show that this method is a special case of the BCR algorithm, from which it immediately follows that this method has the convergence properties stated in Proposition 1.

We describe the dual gradient method below: We fix scalars $\theta_1 \in (0, 1]$ and $\theta_2 \in (0, 1)$. At each iteration, we compute a new dual vector p' from the current dual vector p using the following iteration.

Dual Gradient Iteration.

Given $p \in \Re^n$, choose a vector $u \in \Re^n$ satisfying

$$\langle d(p), u\rangle \geqq \theta_1 \|d(p)\| \cdot \|u\|, \tag{5.12a}$$

and set

$$p' = p + \lambda u, \tag{5.12b}$$

where λ is any positive scalar satisfying

$$0 \leqq \langle d(p + \lambda u), u\rangle \leqq \theta_2 \langle d(p), u\rangle. \tag{5.12c}$$

The stepsize λ can be computed using, say, the Armijo rule [4]. (If Assumption C holds, then exact line search, i.e., $\theta_2 = 0$, is also allowable.)

The above dual gradient method seems to be quite different from the coordinate relaxation methods but, as we show below, it is a special case of the block coordinate relaxation algorithm. (Hence it is convergent in the sense of Proposition 1.)

PROPOSITION 5. *The iteration* (5.12a)-(5.12c) *is well defined and is a special case of the* BCR *iteration for solving* (P^{E}), *with* $\gamma = 1$, $I = N$, $\phi_N(\eta, \pi) = \|\eta\|$, *and* $\delta_N(\eta, \eta') = (1/(1-\theta_2)\theta_1 + 1)\|\eta - \eta'\|$.

Proof. If the iteration (5.12a)-(5.12c) is not well defined, it must be that $\langle d(p+\lambda u), u\rangle$ is bounded away from 0 as $\lambda \to +\infty$. In that case $q(p+\lambda u) \to +\infty$ as $\lambda \to +\infty$, a contradiction of the feasibility of (P^E).

It is clear that p' satisfies (2.2a) and since $\langle d(p'), u\rangle \geqq 0$, p' also satisfies (2.2b) with $\gamma = 1$. To see that p' satisfies (2.2c), note from (5.12a)-(5.12c) that

$$\begin{aligned}\langle d(p)-d(p'), u\rangle &\geqq (1-\theta_2)\langle d(p), u\rangle \\ &\geqq (1-\theta_2)\theta_1 \|d(p)\| \cdot \|u\|.\end{aligned}$$

Hence, by the Cauchy-Schwartz inequality,

$$\|d(p)-d(p')\| \geqq (1-\theta_2)\theta_1 \|d(p)\|.$$

This in turn implies

$$\begin{aligned}\|d(p')\| &\leqq \|d(p)\| + \|d(p)-d(p')\| \\ &\leqq (1/(1-\theta_2)\theta_1 + 1)\|d(p)-d(p')\|.\end{aligned}$$

□

There are a number of choices for the dual ascent direction u. One such choice, proposed in [50] (also see [39, Thm. 4.4.1]), is

$$u = Hd(p), \tag{5.13a}$$

where H is any $n \times n$ symmetric matrix satisfying

$$\mu_1 \|Ey\|^2 \leqq \langle Ey, HEy\rangle \leqq \mu_2 \|Ey\|^2, \quad \forall y \in \Re^m, \tag{5.13b}$$

and μ_1, μ_2 are two positive scalar constants. To see that this choice of u satisfies (5.12a), note that since (P^E) is feasible, there exists $\bar{x} \in \Re^m$ satisfying $b = E\bar{x}$. Hence

$$d(p) = E(\bar{x} - \nabla f^*(E^T p)). \tag{5.14}$$

This, together with (5.13a), (5.13b), implies that

$$\begin{aligned}\langle d(p), u\rangle &= \langle d(p), Hd(p)\rangle \\ &\geqq \mu_1 \|d(p)\|^2 \\ &\geqq \mu_1 \|d(p)\| \, \|u\|/\mu_2,\end{aligned}$$

where the last inequality follows from (cf. (5.13b), (5.14)) $\|Hd(p)\| \leqq \mu_2 \|d(p)\|$. There are a number of choices for the matrix H. For example, H can be computed using a quasi-Newton scheme. Because the matrix E is not assumed to have full row rank, this condition is in general weaker than the requirement that H is positive definite over the entire space.

The identification of the dual gradient method with block coordinate relaxation methods motivates a number of new algorithms. For example, we can consider a block coordinate relaxation algorithm in which each block of coordinates is relaxed either by maximizing the dual functional with respect to all of the coordinates in the block or by performing a one-dimensional line search in the direction given by the partial derivatives of q with respect to these coordinates. The type of relaxation iteration to use for each coordinate block can then be chosen according to the problem structure. As another example, consider the case where the dual functional $q(p)$ has a negative curvature with respect to only a subset of the coordinates of p. In such a case we can take advantage of this structure by alternating between a second-order iteration for these coordinates and a first-order iteration for the remaining coordinates. We illustrate this scheme below.

Suppose that the problem (P^E) has the following separable form

$$\text{Minimize} \quad g(y)+h(z)$$
$$\text{subject to} \quad Ay+Bz=d,$$
$$Cz=e,$$

where $g:\Re^{m_1}\to(-\infty,+\infty]$, $h:\Re^{m_2}\to(-\infty,+\infty]$ are strictly convex functions ($m_1+m_2=m$) and A, B, C, d, e are matrices/vectors of appropriate dimensions. By attaching Lagrange multiplier vectors u and v to, respectively, the constraints $Ay+Bz=d$ and $Cz=e$, we obtain the following dual functional (cf. (1.3)):

$$q(u,v)=\langle d,u\rangle+\langle e,v\rangle-g^*(A^Tu)-h^*(B^Tu+C^Tv),$$

where g^*, h^* denote, respectively, the conjugate function of g and h (cf. (1.2)).

Then, if h^* has a positive curvature, we can use a second-order iteration to update the multiplier vector v. For example, suppose that h is quadratic of the form $h(z)=\langle z,Qz\rangle/2$, where Q is an $m_2\times m_2$ symmetric positive definite matrix. Then $h^*(t)=\langle t,Q^{-1}t\rangle/2$ and

$$\nabla_v q(u,v)=e-CQ^{-1}(B^Tu+C^Tv),$$
$$\nabla_v^2 q(u,v)=-CQ^{-1}C^T,$$

where $\nabla_v q$ and $\nabla_v^2 q$ denote, respectively, the first-order and the second-order partial derivative of q with respect to v. Hence we can update v by moving it along, say, the Newton direction w, given as the solution to the system of linear equations

$$(CQ^{-1}C^T)w=e-CQ^{-1}(B^Tu+C^Tv),$$

to maximize q while u is held fixed. This iteration is a special case of the dual gradient iteration (cf. Proposition 5 and (5.13a), (5.13b)). The other dual vector u can be updated by a different type of BCR iteration.

6. Relation to known methods. In this section, we show that the dual methods proposed in [2], [6], [8]-[10], [13], [15], [18], [21], [23], [24], [27], [28], [31]-[33], [37]-[39], [41], [42], [45], [48]-[50], [57]-[60], [64] are special cases of the BCR algorithm (under the essentially cyclic or the Gauss-Southwell or the quasi-cyclic order of relaxation) and that the conjugate gradient methods in [39], [62] are special cases of the mixed algorithm discussed at the end of § 3. Hence convergence of these methods follow from either Proposition 1 or Proposition 2. We also prove convergence of a general Block S.O.R. algorithm for the solution of the symmetric linear complementarity problem. A mixed version of this algorithm contains as a special case the S.O.R. algorithm proposed in [14].

6.1. General costs and constraints.

PROPOSITION 6. *The methods proposed in* [50] *are special cases of the* BCR *algorithm for solving the special case of* (P^E) *where S is a polyhedral set and f is uniformly convex* (*i.e., satisfies* (4.1) *with* $\omega=2$) *and differentiable on* ri (S).

Proof. The periodic basis ascent method ([50, p. 10]) is a dual single coordinate ascent method that uses exact line search and essentially cyclic order of relaxation. Hence it is a special case of the SCR algorithm with $\alpha_s=0$ for all s. (Although this method allows arbitrary basis vectors to be used for ascent, it can be viewed as a coordinate ascent method, but in a transformed space.) The gradient-type method ([50, p. 11]) is a special case of the gradient method given in § 5.4 where the gradient direction is given by (5.13a), (5.13b) and the line search is exact. (Also see [39], [49] for specialization of this latter method to quadratic programs.) □

PROPOSITION 7. *The methods in* [9] *and* [10] *are special cases of the* SCR *algorithm with* $\alpha_s = 0$ *for all* s.

Proof. Both methods use exact line search. The one in [9] uses cyclic relaxation while the one in [10] uses essentially cyclic relaxation. These methods further require:

(i) S is closed and f is continuously differentiable in ri (S);

(ii) $\{x \in S \mid D(x, y) \leqq \alpha\}$ and $\{y \in \text{ri}\,(S) \mid D(x, y) \leqq \alpha\}$ are bounded for every $y \in \text{ri}\,(S)$ and every $x \in S$, respectively, where $D(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$;

(iii) $\text{argmin}\,\{D(z, y) \mid z \in S,\ E_i z = b_i\} \in \text{ri}\,(S)$, $\forall y \in \text{ri}\,(S)$, $\forall i \in N$.

(The conditions (ii), (iii) do not typically hold, except for special cases such as when f is strongly convex and $S = \Re^m$.)

PROPOSITION 8. *The methods in* [59] *and* [60] *are special cases of the* SCR *algorithm with* $\psi_s(\eta) = \eta$ *for all* s.

Proof. The proof is straightforward from the algorithm description in [59] and § 2 of [60]. In [59], f is further assumed to be separable. (However, the convergence results obtained in [59] are stronger than those obtained from Propositions 1 and 2.) □

6.2. Quadratic costs. In this subsection, we consider the special case of (P) where f is quadratic:

$$f(x) = \langle x, Qx \rangle / 2 + \langle c, x \rangle, \tag{6.1}$$

where Q is an $m \times m$ symmetric positive definite matrix and c is a vector in $\Re^m$. It is easily seen that, under the assumption that (P) is feasible, both Assumptions A and B hold (f is in fact uniformly convex). Direct calculation using (6.1) and (1.2)-(1.5) gives

$$q(p) = -\langle p, Mp \rangle / 2 + \langle w, p \rangle, \tag{6.2}$$

$$\chi(p) = Q^{-1}(E^T p - c), \tag{6.3}$$

$$d(p) = w - Mp, \tag{6.4}$$

where we denote $M = EQ^{-1}E^T$ and $w = b + EQ^{-1}c$.

The first dual coordinate ascent method for solving (6.1) was due to Hildreth [28]. He considered the special case where M is positive definite and proposed a single coordinate cyclic relaxation method with exact line search for its solution. This method was later extended to inexact line search [14], [18], [42], essentially cyclic order of relaxation [27], [38], and block coordinate relaxation [14], [15]. In [14], [27], [38], [42], M is not required to be positive definite.

The general form of this method can be stated as follows: We fix a collection $\mathscr{C}$ of nonempty subsets of N such that their union equals N. We also fix two relaxation parameters ω_1 and ω_2 satisfying $\omega_2 \in (0, 2)$ and $\omega_1 \in (0, \min\{1, \omega_2\}]$. At each iteration, we generate a new dual vector p' from the current dual vector p as follows.

Block S.O.R. *Iteration.*

Given $p \in [0, +\infty)^n$, choose an $I \in \mathscr{C}$. Set

$$p' = (1 - \lambda)p + \lambda \Delta, \tag{6.5}$$

where Δ is any vector in $[0, +\infty)^n$ satisfying

$$\Delta_I = [\Delta_I + w_I - M_I \Delta]^+, \tag{6.6a}$$

$$\Delta_{N \setminus I} = p_{N \setminus I}, \tag{6.6b}$$

and λ is any scalar inside $[\omega_1, \omega_2]$ satisfying $(1 - \lambda)p + \lambda \Delta \geqq 0$.

Note that λ is well defined since $\Delta \geqq 0$. (It can be seen that Δ is equivalently a solution of the problem $\max\{q(\pi) \mid \pi \geqq 0,\ \pi_{N\setminus I} = p_{N\setminus I}\}$.) Below we show that the Block S.O.R. iteration, under certain conditions, is a special case of the BCR iteration.

PROPOSITION 9. *If $\omega_1 = \omega_2 = 1$ or if M_{II} is positive definite for all $I \in \mathscr{C}$, then the Block* S.O.R. *iteration is a special case of the* BCR *iteration, with $\gamma = 2/\omega_2 - 1$, $\phi_I(\eta, \pi) = \sum_{i\in I} |\pi_i - [\pi_i + \eta_i]^+|$, and $\delta_I(\eta, \eta') = \rho \|\eta_I - \eta'_I\|$, where ρ is some positive constant depending on M, ω_1, and ω_2 only.*

Proof. We will show that p' given by (6.5)-(6.6b) satisfies (2.2b), (2.2c) ((2.2a) clearly holds). First we prove that (2.2b) holds. From (6.6a), (6.6b) we have

$$\langle w - M\Delta, \Delta - p\rangle \geqq 0, \tag{6.7}$$

and from (6.2) and (6.5) we have

$$\begin{aligned} \langle p' - p, d(p')\rangle &= \langle p' - p, w - Mp'\rangle \\ &= \langle p' - p, M(\Delta - p')\rangle + \langle p' - p, w - M\Delta\rangle \\ &= (1/\lambda - 1)\langle p' - p, M(p' - p)\rangle + \lambda\langle \Delta - p, w - M\Delta\rangle. \end{aligned} \tag{6.8}$$

Also, (6.2) and (6.4) imply

$$\begin{aligned} q(p') - q(p) &= -\langle p', Mp'\rangle/2 + \langle w, p'\rangle + \langle p, Mp\rangle/2 - \langle w, p\rangle \\ &= \langle p' - p, M(p' - p)\rangle/2 + \langle p' - p, w - Mp'\rangle \\ &= \langle p' - p, M(p' - p)\rangle/2 + \langle p' - p, d(p')\rangle. \end{aligned} \tag{6.9}$$

Multiplying both sides by $2(1/\lambda - 1)$ and using (6.7), (6.8), we obtain

$$\begin{aligned} 2(1/\lambda - 1)[q(p') - q(p)] &= (1/\lambda - 1)\langle p' - p, M(p' - p)\rangle + 2(1/\lambda - 1)\langle p' - p, d(p')\rangle \\ &\leqq \langle p' - p, d(p')\rangle + 2(1/\lambda - 1)\langle p' - p, d(p')\rangle \\ &= (2/\lambda - 1)\langle p' - p, d(p')\rangle. \end{aligned}$$

Since $\lambda \leqq \omega_2$ and $2/\lambda - 1$ is a decreasing function of λ, (2.2b) holds with $\gamma = 2/\omega_2 - 1$.

Now we prove that (2.2c) holds. First suppose that $\omega_1 = \omega_2 = 1$. Then $\lambda = 1$ and it follows from (6.5)-(6.6b) that

$$p'_I = [p'_I + w_I - M_I p']^+.$$

Hence $\phi_I(d(p'), p') = 0$ and (2.2c) holds. Next suppose that M_{II} is positive definite. Denote $\tilde{I} = \{i \in I \mid p'_i + w_i - M_i p' < 0\}$ and $\hat{I} = \{i \in I \mid \Delta_i = 0\}$. Then we have

$$p'_i - [p'_i + w_i - M_i p']^+ = p'_i, \quad \forall i \in \tilde{I}, \tag{6.10}$$

$$p'_i - [p'_i + w_i - M_i p']^+ = M_i p' - w_i, \quad \forall i \in I\setminus\tilde{I}. \tag{6.11}$$

Also, using the fact (cf. (6.6a)) $w_i - M_i\Delta = 0$ for all $i \in I\setminus\hat{I}$, we obtain

$$\begin{aligned} p'_i &= (1 - \lambda)p_i, \quad \forall i \in \hat{I}, \\ w_i - M_i p' &= (1 - \lambda)(w_i - M_i p), \quad \forall i \in I\setminus\hat{I}. \end{aligned}$$

This implies

$$\lambda p'_i = (1 - \lambda)(p_i - p'_i), \quad \forall i \in \hat{I}, \tag{6.12}$$

$$\lambda(w_i - M_i p') = (1 - \lambda)M_i(p' - p), \quad \forall i \in I\setminus\hat{I}, \tag{6.13}$$

which, together with the definition of $\tilde{I}$, implies

$$p'_i \leqq M_i p' - w_i = (1 - 1/\lambda)M_i(p' - p), \quad \forall i \in \tilde{I} \cap (I\setminus\hat{I}), \tag{6.14}$$

$$w_i - M_i p' \geqq -p'_i = (1 - 1/\lambda)(p_i - p'_i), \quad \forall i \in (I\setminus\tilde{I}) \cap \hat{I}. \tag{6.15}$$

Also we have that, for all $i \in I$, $w_i - M_i p' \leqq (1-\lambda)(w_i - M_i p)$ (since $w_i - M_i \Delta \leqq 0$ and $p' = (1-\lambda)p + \lambda\Delta$). Hence

$$\lambda(w_i - M_i p') \leqq (1-\lambda) M_i (p' - p), \quad \forall i \in K, \tag{6.16}$$

where we let $K = \{i \in (I \setminus \tilde{I}) \cap \hat{I} \mid w_i - M_i p' > 0\}$. Combining (6.12)-(6.16), we obtain

$$\begin{aligned}
p_i' &= (1-1/\lambda)(p_i' - p_i), && \forall i \in \tilde{I} \cap \hat{I},\\
p_i' &\leqq (1-1/\lambda) M_i (p' - p), && \forall i \in \tilde{I} \cap (I \setminus \hat{I}),\\
w_i - M_i p' &\geqq (1-1/\lambda)(p_i - p_i'), && \forall i \in ((I \setminus \tilde{I}) \cap \hat{I}) \setminus K,\\
w_i - M_i p' &\leqq (1-1/\lambda) M_i (p - p'), && \forall i \in K,\\
w_i - M_i p' &= (1-1/\lambda) M_i (p - p'), && \forall i \in (I \setminus \tilde{I}) \cap (I \setminus \hat{I}).
\end{aligned}$$

Combining the above with (6.10), (6.11) and using (6.4), we obtain

$$\begin{aligned}
\phi_i(d(p'), p') &\leqq |1/\lambda - 1|\,|p_i - p_i'|, && \forall i \in \hat{I} \setminus K,\\
\phi_i(d(p'), p') &\leqq |1/\lambda - 1|\,|M_i (p - p')|, && \forall i \in I \setminus (\hat{I} \setminus K),
\end{aligned}$$

where $\phi_i(\eta, \pi) = |\pi_i - [\pi_i + \eta_i]^+|$. This, together with (6.6b), implies that

$$\sum_{i \in I} \phi_i(d(p'), p') \leqq |1/\lambda - 1| \rho_1 \| p_I - p_I' \|, \tag{6.17}$$

for some positive constant ρ_1 depending on M_{II} only. Since M_{II} is positive definite,

$$\begin{aligned}
\| p_I - p_I' \|^2 &\leqq \rho_2 \cdot \langle p_I - p_I', M_{II}(p_I - p_I') \rangle\\
&\leqq \rho_2 \cdot \| p_I - p_I' \| \cdot \| M_{II}(p_I - p_I') \|\\
&= \rho_2 \cdot \| p_I - p_I' \| \cdot \| d_I(p) - d_I(p') \|,
\end{aligned}$$

where ρ_2 is some positive constant depending on M_{II} only, and the equality follows from (6.4). This and (6.17) imply that

$$\sum_{i \in I} \phi_i(d(p'), p') \leqq |1/\lambda - 1| \rho_1 \cdot \rho_2 \cdot \| d_I(p) - d_I(p') \|.$$

Since $\lambda \in [\omega_1, \omega_2]$, $|1/\lambda - 1| \leqq \max\{1/\omega_1 - 1, 1 - 1/\omega_2\}$. □

COROLLARY 9. *The methods in* [15], [18], [27], [28], [38], [42] *are special cases of the* BCR *algorithm.*

Proof. The methods in [15], [18], [28] require M to be positive definite, in which case M_{II} is positive definite for any $I \subseteq N$. The methods in [27], [38], [42] use single coordinate relaxation, in which case M_{II} is always positive definite (since E has no zero row). Each of the above methods uses either cyclic or essentially cyclic order of relaxation. □

If M_{II} is not positive definite and $\lambda \neq 1$, then it is possible that $d_I(p) = d_I(p')$ and $p_I' \neq [p_I' + d_I(p')]^+$, in which case there is no continuous δ_I and ϕ_I satisfying, respectively, (2.1a) and (2.1b) for which (2.2c) holds. However, the Block S.O.R. algorithm can still be shown to converge by modifying the proofs in § 3 and § 4. To the best of our knowledge, this is the first proof of convergence for this algorithm that makes no assumption on the problem other than that it be feasible.

PROPOSITION 10. *Let p^r be the iterate generated by the Block* S.O.R. *algorithm at the rth iteration. Then, under either the quasi-cyclic (with $\omega = 2$) or the Gauss-Southwell order of relaxation, $\{\chi(p^r)\} \to x^*$ and $\{q(p^r)\} \to f(x^*)$.*

Proof. From the proof of Proposition 9 and Lemma 2 we have that (3.1a) holds with $\gamma = 2/\omega_2 - 1$. Since (3.1b) and Lemma 1 clearly hold and the proof of Lemma 3(a), (b) depends only on Lemmas 1 and 2, it follows that Lemma 3(a), (b) hold. Suppose that Lemma 3(c) also holds. Then since the proof of Proposition 1(a) depends only on Lemmas 1 and 3 and both these lemmas hold, Proposition 1(a) must hold. Similarly, because the hypothesis of Proposition 1(b) is satisfied (f is everywhere differentiable), Proposition 1(b) also holds. Since f is uniformly convex (i.e., f satisfies (4.1) with $\omega = 2$), an analogous argument shows that parts (a), (b) of Proposition 2 also hold. Therefore it suffices to prove that Lemma 3(c) holds.

Suppose that Lemma 3(c) does not hold. Then there exist a scalar $\varepsilon > 0$, a coordinate block $i \in \mathscr{C}$ and a subsequence $R \subseteq \{1, 2, \cdots\}$ for which the coordinates p_i, $i \in I$, are relaxed at the rth iteration, for all $r \in R$, and

$$\| p_I^r - [p_I^r + d_I(p^r)]^+ \| \geqq \varepsilon, \quad \forall r \in R. \tag{6.18a}$$

Let $x^r = \chi(p^r)$. Since (cf. Lemma 3(a)) $\{x^r\}$ is bounded, by further passing into a subsequence if necessary, we will assume that $\{x^{r-1}\}_{r\in R} \to x^\infty$ for some $x^\infty \in \Re^m$. Let $d = b - Ex^\infty$ and let λ^r, Δ^r denote the λ, Δ generated (cf. (6.5)-(6.6b)) at the rth iteration. Then (cf. (1.4), (6.4))

$$w_I - M_I p^r = d_I(p^r) = b_I - E_I x^r, \quad \forall r. \tag{6.18b}$$

Since $\{x^{r-1}\}_{r\in R} \to x^\infty$, this implies that

$$\{w_I - M_I p^{r-1}\}_{r\in R} \to d_I. \tag{6.19a}$$

Since (cf. (6.5), (6.18b)) $E_I(x^r - x^{r-1}) = M_I(p^r - p^{r-1}) = \lambda^r M_I(\Delta^r - p^{r-1})$ and $\lambda^r \geqq \omega_1 > 0$, it follows from (6.19a) and Lemma 3(b) that

$$\{w_I - M_I \Delta^r\}_{r\in R} \to d_I. \tag{6.19b}$$

Now, by (6.6a), $w_I - M_I \Delta^r \leqq 0$ for all $r \in R$. Hence by (6.19b), $d_I \leqq 0$. Let $I^- = \{i \in I \mid d_i < 0\}$. Then

$$d_i = 0, \quad \forall i \in I \setminus I^-, \tag{6.19c}$$

and (cf. (6.6a), (6.19b)), for all $r \in R$ sufficiently large,

$$\Delta_i^r = 0, \quad \forall i \in I^-. \tag{6.19d}$$

Also, from (6.8) we have that, for all $r \in R$,

$$\langle p^r - p^{r-1}, d(p^r)\rangle = (1/\lambda^r - 1)\langle p^r - p^{r-1}, M(p^r - p^{r-1})\rangle + \lambda^r \langle \Delta^r - p^{r-1}, w - M\Delta^r\rangle,$$

which, together with (6.9), (6.6b), and $\lambda^r \in (0, 2)$, implies that

$$\begin{aligned} q(p^r) - q(p^{r-1}) &= \langle p^r - p^{r-1}, M(p^r - p^{r-1})\rangle/2 + \langle p^r - p^{r-1}, d(p^r)\rangle \\ &= (1/\lambda^r - 1/2)\langle p^r - p^{r-1}, M(p^r - p^{r-1})\rangle + \lambda^r \langle \Delta^r - p^{r-1}, w - M\Delta^r\rangle \\ &\geqq \lambda^r \langle \Delta_I^r - p_I^{r-1}, w_I - M_I \Delta^r\rangle. \end{aligned}$$

Since $\lambda^r \geqq \omega_1 > 0$ for all r and the right-hand side of the above is nonnegative by (6.6a), we have $\{\langle \Delta_I^r - p_I^{r-1}, w_I - M_I \Delta^r\rangle\}_{r\in R} \to 0$. This, together with (6.19a)-(6.19d), implies that

$$\{p_i^r\}_{r\in R} \to 0, \quad \forall i \in I^-,$$

$$\{w_i - M_i p^r\}_{r\in R} \to 0, \quad \forall i \in I \setminus I^-.$$

Hence $\{p_I^r - [p_I^r + d_I(p^r)]^+\}_{r\in R} \to 0$, a contradiction of (6.18a). □

Analogous to the mixed algorithm discussed at the end of § 3, we can consider an algorithm whereby other dual ascent iterations are inserted between the Block S.O.R. iterations at regular intervals. It can be shown that Proposition 10 also holds for this mixed algorithm, provided that the inserted dual ascent iterations satisfy the condition (3.9). An interesting special case of this mixed algorithm is the Cottle-Pang algorithm proposed in [14]. This algorithm can be seen to be a special case of the Block S.O.R. algorithm using $\omega_1 = \min\{1, \omega_2\}$ and cyclic order of relaxation. The only difference is that a "reduction" step is inserted at the end of each cycle. This reduction step generates a new iterate p' from the current iterate p by the formula:

$$p' = p - \theta u, \tag{6.20}$$

where u is some nonnegative vector in $\Re^n$ satisfying

$$E^T u = 0, \qquad \langle b, u\rangle = 0, \tag{6.21}$$

and θ is the largest scalar for which p' given by (6.20) is nonnegative (if $u = 0$, θ is set to 0). From (6.20), (6.21) we see that $E^T p = E^T p'$ and $\langle b, p\rangle = \langle b, p'\rangle$. Therefore (cf. (1.3) and (1.5))

$$q(p) = q(p'), \qquad \chi(p) = \chi(p'),$$

and the iteration (6.20), (6.21) satisfies (3.9).

Notes and extensions.

1. For each $I \subseteq N$, the matrix M_{II} is positive definite if and only if E_I has full row rank.

2. The vector Δ satisfying (6.6a), (6.6b) can be computed either approximately using iterative methods [39], [43] or exactly using direct methods [16], [17], [36].

3. If the dual functional q given by (6.2) has bounded level sets on the nonnegative orthant in $\Re^n$ (which can be seen to hold if and only if there exists $z \in \Re^n$ such that $Mz - w > 0$), then the sequence of iterates $\{p^r\}$ generated by the Block S.O.R. iteration is bounded and each of its limit points is an optimal dual solution. If E does not have full row rank, the technique discussed at the end of § 3 may be used to maintain $\{p^r\}$ to be bounded. On the other hand, if $\omega_2 \in (0, 1]$ and b lies in the column space of E, the Block S.O.R. iteration can be implemented working with $E^T p$ instead of p. By (6.3) and Lemma 3(a), $\{E^T p^r\}$ is bounded.

4. Consider the special case of (P^E) where f is the sum of a strictly convex quadratic function and the indicator function for a polyhedral set in $\Re^m$. In [39, § 4] a conjugate gradient method with periodic restart (with a gradient iteration as the spacer step) was proposed to solve this problem, and convergence for this method was established for the Polak-Ribiere-Polyak [4] formula. (Also see [62] for computational results on network flow problems.) However, because each conjugate gradient iteration maximizes exactly the dual functional q along some direction, this method is simply a special case of the mixed algorithm discussed at the end of § 3. Hence it immediately follows that this method converges for any conjugate gradient formula and for the general problem (P^E). Furthermore, instead of the gradient iteration, we can use any BCR iteration as the spacer step.

6.3. Entropy costs. In this subsection we consider the following special case of (P^E):

$$\begin{aligned} &\text{Minimize} && f_1(x) = \sum_j x_j \ln (x_j / u_j) \\ &\text{subject to} && Ex = b,\ x \geqq 0, \end{aligned} \tag{6.22}$$

where E is an $n \times m$ matrix, b is a vector in $\Re^n$, and the u_j's are given positive constants. (Here $\ln(\cdot)$ denotes the natural logarithm.) It is easily verified that Assumptions A and B hold for this problem (assuming that it is feasible). This problem, called the entropy maximization problem ($-f_1$ is the classical entropy function weighted by the u_j's), has applications in matrix balancing [2], [8], [21], [24], [31]-[33], [41], [48], [57], ([5, § 5.5.4]), image reconstruction [11], [12], [23], [37], [53] and maximum likelihood estimation [19].

PROPOSITION 11. *The matrix balancing methods in* [2], [8], [21], [24], [31]-[33], [41], [48], [57] *are special cases of the* SCR *algorithm for solving* (6.22) *with* $\alpha_s = 0$ *for all* s.

Proof. In [33] it was shown that the matrix balancing methods in [1], [21], [24], [31], [32], [41], [48], [57] are special cases of Bregman's method [9], [10] for solving (6.22). The RAS-algorithm considered in [2], [8] can also be seen to be such a special case. (The modified RAS-algorithm in [2] uses essentially cyclic instead of cyclic order of relaxation.) Therefore, by Proposition 7, they are special cases of the SCR algorithm with $\alpha_s = 0$ for all s. □

Consider the following special case of (6.22):

$$\begin{aligned} &\text{Minimize} \quad f_2(x) = \sum_j x_j \ln(x_j) \\ &\text{subject to} \quad Ex = b, \qquad x \geqq 0, \end{aligned} \tag{6.23}$$

where (for each i) $b_i > 0$, $e_{ij} \in [0, 1]$ for all j, and $e_{ij} > 0$ for at least one j. The following method for solving (6.23) was proposed in [23] and [37]. It begins with any $x \in \Re^m$ satisfying $x_j = \exp(\sum_i e_{ij} p_i - 1)$, for all j, for some $p \in \Re^n$. (Here $\exp(\cdot)$ denotes the exponential function.) Given an $x \in \Re^m$, it generates a new estimate x' as follows.

Multiplicative ART *Iteration.*

Choose an index $s \in N$ and set

$$x_j' = x_j \left(b_s \Big/ \left(\sum_k e_{sk} x_k \right) \right)^{e_{sj}}, \quad \forall j = 1, \cdots, m. \tag{6.24}$$

(The index s is chosen by the essentially cyclic order.) The iteration (6.24) is also a special case of the SCR iteration, as we show below.

PROPOSITION 12. *The multiplicative* ART *method is a special case of the* SCR *algorithm with* $\alpha_i = 1 - \min_j \{e_{ij} \mid e_{ij} > 0\}$ *and*

$$\psi_i(\eta) = \begin{cases} \ln(b_i/(b_i - \eta)) & \text{if } b_i - \eta > 0, \\ -\infty & \text{otherwise.} \end{cases}$$

Proof. Straightforward calculation finds the conjugate function of f_2 to be $\sum_j g_j(t_j)$, where $g_j(t_j) = \exp(t_j - 1)$. Hence $\nabla g_j(t_j) = \exp(t_j - 1)$ and (cf. (1.4) and (1.5))

$$\chi_j(p) = \exp(t_j - 1), \tag{6.25a}$$

$$b_i - d_i(p) = \sum_j e_{ij} \exp(t_j - 1), \tag{6.25b}$$

where $t_j = \sum_i e_{ij} p_i$ and $\chi_j(p)$ denotes the jth coordinate of $\chi(p)$.

Given $p \in \Re^n$ and $s \in N$, consider the single coordinate relaxation

$$p' = p + \lambda e^s, \tag{6.26}$$

where e^s denotes the sth coordinate vector in $\Re^n$ and λ is the scalar satisfying

$$b_s/(b_s - d_s(p)) = \exp(\lambda). \tag{6.27}$$

Then (cf. (6.25a))

$$\begin{aligned}\chi_j(p') &= \chi_j(p+\lambda\, e^s)\\ &= \exp\left(\sum_i e_{ij}p_i + e_{sj}\lambda - 1\right)\\ &= \chi_j(p)\exp(\lambda)^{e_{sj}}\\ &= \chi_j(p)(b_s/(b_s-d_s(p)))^{e_{sj}}.\end{aligned}$$

Comparing the above equation with (6.24), we see that the two iterations (6.24) and (6.26), (6.27) are equivalent.

We now show that p' generated by the iteration (6.26), (6.27) satisfies (5.11a)-(5.11c). For simplicity we assume that $d_s(p)\geqq 0$ (the case where $d_s(p)<0$ can be treated analogously). Then, by (6.27), $\exp(\lambda)\geqq 1$. Since (cf. (6.25b))

$$\begin{aligned}b_s - d_s(p') &= b_s - d_s(p+\lambda\, e^s)\\ &= \sum_j e_{sj}\cdot\exp(t_j + e_{sj}\lambda - 1)\\ &= \sum_j e_{sj}\cdot\exp(t_j-1)\cdot\exp(\lambda)^{e_{sj}},\end{aligned}\tag{6.28}$$

we obtain from the facts $\exp(\lambda)\geqq 1$ and $e_{sj}\geqq 1-\alpha_s$ that

$$\begin{aligned}b_s - d_s(p') &\geqq \sum_j e_{sj}\cdot\exp(t_j-1)\cdot\exp(\lambda)^{1-\alpha_s}\\ &= (b_s-d_s(p))\cdot\exp(\lambda)^{1-\alpha_s}.\end{aligned}$$

This, together with (6.27), implies that $1-d_s(p')/b_s \geqq (1-d_s(p)/b_s)^{\alpha_s}$, or, equivalently,

$$\psi_s(d_s(p'))\leqq \alpha_s\psi_s(d_s(p)).$$

On the other hand, since $e_{sj}\in[0,1]$ for all j, we have from (6.28) and the fact $\exp(\lambda)>1$,

$$\begin{aligned}b_s - d_s(p') &= \sum_j e_{sj}\exp(t_j-1)\cdot\exp(\lambda)^{e_{sj}}\\ &\leqq \sum_j e_{sj}\exp(t_j-1)\cdot\exp(\lambda)\\ &= (b_s-d_s(p))\exp(\lambda) = b_s.\end{aligned}$$

Hence $\psi_s(d_s(p'))\geqq 0$ so that λ satisfies (5.11a). □

There is, however, a slight difficulty with the choice of ψ_i given above, namely, that ψ_i is not continuous everywhere. This difficulty can be circumvented by redefining ψ_i on the interval $[b_i-\varepsilon, +\infty)$ to be a continuous (and strictly increasing) extension of itself, for some $\varepsilon>0$. For this choice of ψ_i, the proof of Proposition 12 still goes through, provided that it holds

$$d_{s^r}(p^r)\leqq b_{s^r}-\varepsilon,\quad d_{s^r}(p^{r-1})\leqq b_{s^r}-\varepsilon,\quad \forall r=0,1,\cdots,$$

where p^r denotes the iterate generated by (6.26), (6.27) at the rth iteration and s^r is the index of the coordinate relaxed at the rth iteration. (Hence the value of ε depends on p^0.) To see that this indeed holds, let q denote the dual functional (cf. (1.3)) given by

$$\begin{aligned}q(p) &= \min_{x\geqq 0}\{f_2(x)+\langle p, b-Ex\rangle\}\\ &= \langle b,p\rangle - \sum_j \exp\left(\sum_i e_{ij}p_i - 1\right),\quad \forall p\in\Re^n.\end{aligned}$$

Hence, for any $p \in \Re^n$, any $s \in N$, and any λ given by (6.27), we have

$$\begin{aligned} q(p+\lambda e^s) - q(p) &= b_s\lambda + \sum_j \exp\left(\sum_i e_{ij}p_i - 1\right)(1-\exp(e_{sj}\lambda)) \\ &= b_s\lambda + \sum_j \chi_j(p)(1-\exp(e_{sj}\lambda)) \\ &\geqq b_s\lambda - \sum_j \chi_j(p)e_{sj}(\exp(\lambda)-1) \\ &= b_s(\lambda - 1 + \exp(-\lambda)), \end{aligned}$$

where the second equality follows from (6.25a) and the inequality follows from the fact $\omega^\theta \leqq 1+\theta(\omega-1)$ for all $\omega \geqq 0$ and all $\theta \in [0,1]$. Since the function $\eta \to \eta - 1 + \exp(-\eta)$ is nonnegative for all positive η and attains the value zero at $\eta = 0$ only, this, together with the observation that $\{q(p^r) - q(p^{r-1})\} \to 0$ and $b_s > 0$ for all s, implies (cf. (6.26), (6.27))

$$\{p^r - p^{r-1}\} \to 0, \qquad \{(b_{s^r} - d_{s^r}(p^{r-1}))/b_{s^r}\} \to 1.$$

It then follows from (6.25a), (6.25b) that $\{d(p^r) - d(p^{r-1})\} \to 0$ and $\{d_{s^r}(p^r)/b_{s^r}\} \to 0$. Hence both $\{b_{s^r} - d_{s^r}(p^r)\}$ and $\{b_{s^r} - d_{s^r}(p^{r-1})\}$ are bounded away from zero. (The above argument is based on one given in [37] for Lemma 1 therein.)

6.4. Network flow constraints. In this subsection we consider the special case of (P^E) where E is the node-arc incidence matrix for a generalized network (i.e., each column of E has at most one positive entry and at most one negative entry). An important special case of this problem is the pure network flow problem, for which each positive entry is $+1$ and each negative entry is -1.

PROPOSITION 13. *The network flow methods in* [13], [45], [58], [64] *are special cases of the* SCR *algorithm with* $\alpha_s = 0$ *for all* s.

Proof. These methods are all dual single coordinate relaxation algorithms that use exact line search and cyclic order of relaxation. In [13], [45], [64], the cost function f is further assumed to be separable. In [58] the cost function f is assumed to be strongly convex. (The references [45], [64] do not contain convergence proofs. In [58], to prove convergence, it is also assumed that an optimal dual solution exists and is unique, and that the dual functional q is *twice* differentiable.) □

PROPOSITION 14. *The network flow method in* § 2 *of* [6] *is a special case of the* SCR *algorithm with* $\psi_s(\eta) = \eta$ *for all* s.

Proof. The proof is straightforward from the algorithm description in § 2 of [6]. The cost function f is further assumed to be separable. (However, [6] allows *arbitrary* order of relaxation and further establishes convergence to an optimal dual solution (assuming only that an optimal dual solution exists).) □

Note. By applying the results in §§ 4 and 5, we can readily extend many of the methods discussed in this section. As an example, since (cf. Lemma 3(a)) the sequence of primal vectors generated by the BCR iteration remains in a compact subset of S and the entropy cost f_1 is strongly convex in any compact subset of S, the methods described in [2], [8], [21], [23], [24], [31]-[33], [37], [41], [48], [57] can also be implemented using the quasi-cyclic order of relaxation.

7. A parallel line search procedure. In this section, we present a technique for parallelizing the inexact line search step in the SCR iteration when f has a certain separable structure. This technique is most suited for problems where the constraint matrix E has relatively sparse rows.

Suppose that f is *block separable*, in the sense that

$$f(x) = \sum_{J \in \mathcal{D}} f_J(x_J), \tag{7.1}$$

where $\mathcal{D}$ is a collection of nonempty, pairwise disjoint subsets of $M = \{1, \cdots, m\}$ and each $f_J : \Re^{|J|} \to (-\infty, +\infty]$ is strictly convex function ($\mathcal{D} = \{M\}$ is a valid, but uninteresting choice). We will show that the stepsize λ in the SCR iteration, with $\psi_s(\eta) = \eta$, can be calculated in parallel using at most $|\mathcal{D}|$ processors. (Extensions to arbitrary ψ_s and to the BCR iteration are straightforward, but for simplicity we will not consider them here.)

Denote by f_J^* the conjugate function of f_J and, for each $i \in N$, denote

$$\mathcal{D}(i) = \{J \in \mathcal{D} \mid e_{ij} \neq 0 \text{ for some } j \in J\}.$$

For each $i \in N$, let $\{\rho_{iJ}\}_{J \in \mathcal{D}(i)}$ be any set of positive scalars satisfying

$$\sum_{J \in \mathcal{D}(i)} \rho_{iJ} = 1.$$

Let μ be any scalar in the interval $(0, 1)$. For any nonnegative $p \in \Re^n$ and $s \in N$ satisfying $\beta = d_s(p) \geqq 0$ (the case where $\beta < 0$ may be treated analogously), consider the following procedure for computing an inexact line search stepsize λ.

1. For each $J \in \mathcal{D}(s)$, let $h_J : [0, +\infty) \to [0, +\infty)$ be the function $h_J(\theta) = E_{sJ}(\nabla f_J^*(t_J + \theta(E_{sJ})^T) - \nabla f_J^*(t_J))$, where $t_J = (E_{NJ})^T p$. If $h_J(\theta) \leqq \mu\beta\rho_{sJ}$ for all $\theta \geqq 0$, set $\lambda_J = +\infty$; otherwise, compute a λ_J satisfying

$$\mu\beta\rho_{sJ} \leqq h_J(\lambda_J) \leqq \beta\rho_{sJ}. \tag{7.2}$$

2. Set

$$\lambda = \min_{J \in \mathcal{D}(s)} \{\lambda_J\}. \tag{7.3}$$

(If Assumption C holds, then $\mu = 1$ (i.e., exact line search) is also allowable.) Each step in the above procedure can be seen to be parallelizable among $|\mathcal{D}(s)|$ processors. We have the following main result.

PROPOSITION 15. *For any $p \in [0, +\infty)^n$ and any $s \in N$ such that $\beta = d_s(p) > 0$, the scalar λ given by* (7.2), (7.3) *satisfies*

$$(1 - \mu \min_{J \in \mathcal{D}(s)} \{\rho_{sJ}\})\beta \geqq d_s(p + \lambda e^s) \geqq 0. \tag{7.4}$$

Proof. Since f satisfies (7.1), we obtain from (1.2) that $f^*(t) = \sum_{J \in \mathcal{D}} f_J^*(t_J)$ for all $t \in \Re^m$. Hence (cf. (1.4), (1.5))

$$d_s(p + \theta e^s) = b_s - \sum_{J \in \mathcal{D}(s)} E_{sJ} \nabla f_J^*(t_J + \theta(E_{sJ})^T), \quad \forall \theta \in \Re, \tag{7.5}$$

where $t_J = (E_{NJ})^T p$.

We claim that each λ_J is positive and $\lambda < +\infty$. Each λ_J is positive because $h_J(0) = 0$ and (by convexity of f_J^*) $h_J(\theta)$ is monotonically increasing with θ. To see that $\lambda < +\infty$, suppose the contrary. Then it must be true that

$$E_{sJ}(\nabla f_J^*(t_J + \theta(E_{sJ})^T) - \nabla f_J^*(t_J)) = h_J(\theta) < \mu\beta\rho_{sJ}, \quad \forall \theta \geqq 0,$$

for all $J \in \mathcal{D}(s)$. By summing the above inequality over all $J \in \mathcal{D}(s)$ and using (7.5) and $\sum_{J \in \mathcal{D}(s)} \rho_{sJ} = 1$, we obtain

$$-d_s(p + \theta e^s) + \beta < \mu\beta, \quad \forall \theta \geqq 0,$$

or, equivalently,

$$d_s(p + \theta e^s) > \beta(1 - \mu) > 0, \quad \forall \theta \geqq 0.$$

Hence $q(p+\theta e^s)\to+\infty$ as $\theta\to+\infty$, a contradiction of the assumption that (P) is feasible.

Now we prove (7.4). To prove the second inequality in (7.4), note that (cf. (7.2), (7.3), and the fact that h_J is an increasing function)

$$h_J(\lambda)\leqq\beta\rho_{sJ},\quad\forall J\in\mathcal{D}(s),$$

from which it follows that

$$d_s(p+\lambda e^s)=\beta-\sum_{J\in\mathcal{D}(s)}h_J(\lambda)\geqq 0.$$

To prove the first inequality in (7.4), note that since $\lambda<+\infty$, there exists some $\bar{J}\in\mathcal{D}(s)$ for which

$$\mu\beta\rho_{s\bar{J}}\leqq h_{\bar{J}}(\lambda).$$

Since (cf. $\lambda>0$, $h_J(0)=0$, and $h_J(\theta)$ increases with θ) $h_J(\lambda)\geqq 0$ for all $J\in\mathcal{D}(s)$, this implies that

$$\begin{aligned}d_s(p+\lambda e^s)&=\beta-\sum_{J\in\mathcal{D}(s)}h_J(\lambda)\\&\leqq\beta-\mu\beta\rho_{s\bar{J}}.\end{aligned}$$

□

From Proposition 15 and (5.11a) we see that, for the case $\beta\geqq 0$, the procedure (7.2), (7.3) implements the SCR iteration with $\psi_s(\eta)=\eta$ and $\alpha_s=1-\mu\min_{J\in\mathcal{D}(s)}\{\rho_{sJ}\}$.

To illustrate one computational advantage of the procedure (7.2), (7.3), suppose that $\mathcal{D}=\{\{1\},\cdots,\{m\}\}$ and $f_j^*(t_j)=c_j\exp(t_j)$, where each c_j is a positive scalar. Then, instead of computing λ as an approximate zero of $h(\theta)=\sum_{j\in\mathcal{D}(s)}e_{sj}c_j\exp(t_j+\theta e_{sj})$, say, using an iterative method, we simply set (again assuming that $\beta>0$) $\lambda=\min_{j\in\mathcal{D}(s)}\{\lambda_j\}$, where

$$\lambda_j=(\ln(\exp(t_j)-\mu\beta\rho_{sj}/(e_{sj}c_j)-t_j)/e_{sj},$$

if the quantity inside the ln is positive and $\lambda_j=+\infty$ otherwise.

The line search procedure (7.2), (7.3) is particularly well suited for implementation on fine-grained SIMD (Single Instruction Multiple Data) multiprocessors such as the Connection Machine [29]. To see this, note that this procedure uses the *same* sequence of calculations for all coordinates (this is not true for exact line search). Hence we can execute this procedure *simultaneously* for any subset of coordinates. (In particular, we can choose these coordinates to be pairwise uncoupled.) The communication cost for this procedure depends on the architecture of the machine. On the Connection Machine, by assigning $|\mathcal{D}(s)|$ processors to each $s\in N$ and using an implementation technique similar to that used in [65] for network flow problems, the communication cost can be kept very low.

A potential disadvantage of the procedure (7.2), (7.3) is that the stepsize that it generates is too conservative. Initial computational tests on quadratic cost network flow problems suggest that this is not the case, provided that the row of E operated on is relatively sparse. For dense rows, we can use either exact line search or some refinement of the above procedure. For example, we can dynamically adjust the weights $\{\rho_{sJ}\}$ by giving higher values to those ρ_{sJ} for which h_J has a high growth rate near 0 (ideally the λ_J's would be equal). An interesting research topic is that of finding efficient algorithms for performing this adjustment. Also, instead of choosing λ to be the smallest of the λ_J's, we can choose λ to be the λ_J that yields the largest improvement in the dual cost.

8. Conclusion and extensions. In this paper, we have presented a general algorithmic framework for dual coordinate/gradient ascent and have unified a number of existing methods under this framework. There are many directions in which our results can be extended. For example, we can use a *linear convex combination* of directions generated by the BCR iteration for dual ascent. Such an approach was proposed in [11], [12], [30] in the special cases of single coordinate relaxation for quadratic programming and for entropy maximization, but it also applies to the more general case of the BCR iteration for problems where the cost function is differentiable in the relative interior of its effective domain. Alternatively, we can study specialization of the BCR algorithm to special cases that exploit the structure of the problem. For example, can the conjugate gradient method be efficiently adapted for entropy maximization? It is also worthwhile to implement some of these algorithms (on either a sequential or a parallel machine) in order to test their practical efficiency.

REFERENCES

[1] A. BACHEM AND B. KORTE, *An algorithm for quadratic optimization over transportation polytopes*, Zeitschrift fur Angewante Mathematik and Mechanik, 58 (1978), pp. 459-461.

[2] ———, *On the* RAS-*algorithm*, Computing, 23 (1979), pp. 189-198.

[3] ———, *Minimum norm problems over transportation polytopes*, Linear Algebra Appl., 31 (1980), pp. 103-118.

[4] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, NY, 1982.

[5] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[6] D. P. BERTSEKAS, P. A. HOSEIN, AND P. TSENG, *Relaxation methods for network flow problems with convex arc costs*, SIAM J. Control Optim., 25 (1987), pp. 1219-1243.

[7] G. R. BITRAN AND A. C. HAX, *Disaggregation and resource allocation using convex knapsack problems with bounded variables*, Management. Sci., 27 (1981), pp. 431-441.

[8] L. M. BREGMAN, *Proof of convergence of Sheleikhovskii's method for a problem with transportation constraints*, USSR Comput. Math. and Math. Phys., 1 (1967), pp. 191-204.

[9] ———, *The relaxation method of finding the common point convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. and Math. Phys., 7 (1967), pp. 200-217.

[10] Y. CENSOR AND A. LENT, *An iterative row-action method for interval convex programming*, J. Optim. Theory Appl., 34 (1981), pp. 321-352.

[11] Y. CENSOR AND S. SEGMAN, *On block-iterative entropy maximization*, J. Inform. Optim. Sci., 8 (1987), pp. 275-291.

[12] Y. CENSOR, *Parallel application of block-iterative methods in medical imaging and radiation therapy*, Math. Programming, Series B, 42 (1988), pp. 307-325.

[13] R. W. COTTLE, S. G. DUVALL, AND K. ZIKAN, *A Lagrangian relaxation algorithm for the constrained matrix problem*, Naval Res. Logist. Quart., 33 (1986), pp. 55-76.

[14] R. W. COTTLE AND J. S. PANG, *On the convergence of a block successive over-relaxation method for a class of linear complementarity problems*, Math. Programming Stud., 17 (1982), pp. 126-138.

[15] R. W. COTTLE, G. H. GOLUB, AND R. S. SACHER, *On the solution of large, structured linear complementarity problems: the block partitioned case*, J. Appl. Math. Optim., 4 (1978), pp. 347-363.

[16] R. W. COTTLE AND R. S. SACHER, *On the solution of large, structured, linear complementarity problems: the tridiagonal case*, J. Appl. Math. Optim., 3 (1977), pp. 321-340.

[17] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory of Math. Prog.*, Linear Algebra Appl., 1 (1968), pp. 103-125.

[18] C. W. CRYER, *The solution of a quadratic programming problem using systematic overrelaxation*, SIAM J. Control Optim., 9 (1971), pp. 385-392.

[19] J. N. DARROCH AND D. RATCLIFF, *Generalized iterative scaling for log-linear models*, Ann. Math. Statist., 43 (1972), pp. 1470-1480.

[20] D. A. D'ESOPO, *A convex programming procedure*, Naval Res. Logist. Quart., 6 (1959), pp. 33-42.

[21] S. P. EVANS AND H. R. KIRBY, *A three-dimensional Furness procedure for calibrating gravity models*, Transportation Res., 8 (1974), pp. 105-122.

[22] EVERETT, *Generalized Lagrange multiplier method for solving problems of optimum allocation of resources*, Oper. Res., 11 (1963), pp. 399-417.

[23] R. GORDON, R. BENDER, AND G. T. HERMAN, *Algebraic reconstruction techniques* (ART) *for three-dimensional electron microscopy and X-ray photography*, J. Theoret. Biol., 29 (1970), pp. 471-481.

[24] J. GRAD, *Matrix balancing*, Comput. J., 14 (1971), pp. 280-284.

[25] C. D. HA, *An algorithm for structured, large-scale quadratic programming problems*, Tech. Report 2276, Mathematics Research Center, University of Wisconsin, Madison, 1980.

[26] R. V. HELGASON, J. L. KENNINGTON, AND H. LALL, *Polynomially bounded algorithm for a single constrained quadratic program*, Math. Programming, 18 (1980), pp. 338-343.

[27] G. T. HERMAN AND A. LENT, *A family of iterative quadratic optimization algorithms for pairs of inequalities, with application in diagnostic radiology*, Math. Programming Stud., 9 (1978), pp. 15-29.

[28] C. HILDRETH, *A quadratic programming procedure*, Naval Res. Logist. Quart., 4 (1957), pp. 79-85; see also *Erratum*, Naval Res. Logist. Quart., 4 (1957), p. 361.

[29] W. D. HILLIS, *The Connection Machine*, M.I.T. Press, Cambridge, MA, 1985.

[30] A. N. IUSEM AND A. R. DE PIERRO, *A simultaneous iterative method for computing projections on polyhedra*, SIAM J. Control Optim., 25 (1987), pp. 231-243.

[31] T. R. JEFFERSON AND C. H. SCOTT, *The analysis of entropy models with equality and inequality constraints*, Transportation Res., 138 (1979), pp. 123-132.

[32] J. KRUITHOF, *Calculation of telephone traffic*, De Ingenieur (E. Elektrotechnik 3), 52 (1937), pp. E15-E25.

[33] B. LAMOND AND N. F. STEWART, *Bregman's balancing method*, Transportation Res.-B, 15B (1981), pp. 239-248.

[34] L. S. LASDON, *Optimization Theory for Large Systems*, MacMillan, New York, NY, 1970.

[35] C. LEMARECHAL AND R. MIFFLIN, *Global and superlinear convergence of an algorithm for one dimensional minimization of convex functions*, Math. Programming, 24 (1982), pp. 241-256.

[36] C. E. LEMKE, *On complementary pivot theory*, in Mathematics of the Decision Sciences, Part I, G. B. Dantzig and A. F. Veinott, Jr., eds., American Mathematical Society, Providence, RI, 1968.

[37] A. LENT, *A convergent algorithm for maximum entropy image restoration with a medical X-ray application*, in Image Analysis and Evaluation, R. Shaw, ed., Society of Photographic Scientists and Engineers (SPSE), Washington, D.C., 1977, pp. 249-257.

[38] A. LENT AND Y. CENSOR, *Extensions of Hildreth's row-action method for quadratic programming*, SIAM J. Control Optim., 18 (1980), pp. 444-454.

[39] Y. Y. LIN AND J. S. PANG, *Iterative methods for large convex quadratic programs: a survey*, SIAM J. Control Optim., 25 (1987), pp. 383-411.

[40] D. G. LUENBERGER, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[41] S. H. MACGILL, *Convergence and related properties for a modified biproportional problem*, Environ. Plan A, 11 (1979), pp. 499-506.

[42] O. L. MANGASARIAN, *Sparsity-preserving* SOR *algorithms for separable quadratic and linear programming*, Comput. Oper. Res., 11 (1984), pp. 105-112.

[43] O. L. MANGASARIAN AND R. DE LEONE, *Parallel gradient projection successive overrelaxation for symmetric linear complementarity problems and linear programs*, in Ann. Oper. Res. 14: Parallel Optimization on Novel Computer Architectures, R. R. Meyer and S. A. Zenios, eds., Baltzer, Switzerland, 1988, pp. 41-59.

[44] R. MIFFLIN, *An implementation of an algorithm for univariate minimization and an application to nested optimization*, Math. Programming Stud., 31 (1987), pp. 155-166.

[45] A. OHUCHI AND I. KAJI, *Lagrangian dual coordinatewise maximization algorithm for network transportation problems with quadratic costs*, Networks, 14 (1984), pp. 515-530.

[46] ———, *Algorithms for optimal allocation problems having quadratic objective functions*, J. Oper. Res. Soc. Japan, 23 (1980), pp. 64-80.

[47] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, NY, 1970.

[48] E. E. OSBORNE, *On pre-conditioning of matrices*, J. Assoc. Comput. Mach., 7 (1960), pp. 338-345.

[49] J. S. PANG, *More results on the convergence of iterative methods for the symmetric linear complementarity problem*, J. Optim. Theory Appl., 49 (1986), pp. 107-134.

[50] ———, *On the convergence of dual ascent methods for large scale linearly constrained optimization problems*, School of Management, The University of Texas at Dallas, TX, June 1984.

[51] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, NY, 1971.

[52] M. J. D. POWELL, *On search directions for minimization algorithms*, Math. Programming, 4 (1973), pp. 193–201.

[53] ———, *An algorithm for maximizing entropy subject to simple bounds*, Math. Programming, 42 (1988), pp. 171–180.

[54] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[55] ———, *Network Flows and Monotropic Optimization*, Wiley-Interscience, New York, NY, 1984.

[56] R. W. H. SARGENT AND D. J. SEBASTIAN, *On the convergence of sequential minimization algorithms*, J. Optim. Theory Appl., 12 (1973), pp. 567–575.

[57] R. SINKHORN, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, Ann. Math. Statist., 35 (1964), pp. 876–879.

[58] T. E. STERN, *A class of decentralized routing algorithms using relaxation*, IEEE Trans. Comm., COM-25 (1977), pp. 1092–1102.

[59] P. TSENG AND D. P. BERTSEKAS, *Relaxation methods for problems with strictly convex separable costs and linear constraints*, Math. Programming, 38 (1987), pp. 303–321.

[60] ———, *Relaxation methods for problems with strictly convex costs and linear inequality constraints*, LIDS-P-1717, Laboratory for Information and Decision Systems, M.I.T., MA, 1987; Math. Oper. Res., to appear.

[61] H. UZAWA, *Iterative methods for concave programming*, in Studies in Linear and Non-linear Programming, K. J. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958.

[62] J. A. VENTURA AND D. W. HEARN, *Computational development of a Lagrangian dual approach for quadratic networks*, Report 88-14, Industrial and Systems Engineering Department, University of Florida, FL, December 1988.

[63] W. I. ZANGWILL, *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[64] S. A. ZENIOS AND J. M. MULVEY, *Relaxation techniques for strictly convex network problems*, in Ann. Oper. Res. 5: Algorithms and Software for Optimization, C. L. Monma, ed., Baltzer, Switzerland, 1986, pp. 517–538.

[65] S. A. ZENIOS AND R. LASKEN, *Nonlinear network optimization on a massively parallel connection Machine*, in Ann. Oper. Res. 14: Parallel Optimization on Novel Computer Architecture, R. R. Meyer and S. A. Zenios, eds., Baltzer, Switzerland, 1988, pp. 147–165.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 243–249, January 1990

012

ERRATUM: Optimal Control of Strongly Monotone Variational Inequalities*

SHUZHONG SHI†

I thank Professor V. Barbu for showing me that in my paper entitled "Optimal Control of Strongly Monotone Variational Inequalities" [*SIAM J. Control Optim.*, 26 (1988), pp. 274–290], the inequality (4.10) is wrong (except $K = V$) and the second relation of (4.5) is false. This is a serious error; in fact, I must replace $N_K(\bar{y})$ in (4.5) by a significant "directional derivative" of $N_K(\cdot)$ as in Theorem 3.1 of Barbu's book [2], which, unfortunately, I did not read well enough before.

The corrected second relation of (4.5) is as follows:

$$F'(\bar{y})^*\bar{p} \in -D_P N_K(\bar{y}, -F(\bar{y}) - E(\bar{u}))(\bar{p}) + \partial g(\bar{y}) \tag{1}$$

where $N_K(\cdot)$ is the normal cone to K that, as a set-valued map from K to V^*, has a closed graph for the norm topology of K and the weak topology of V^* and where, for $(y, y^*) \in \text{graph}\,(N)$:

$$\begin{aligned} &q^* \in D_P N_K(y, y^*)(p) \Leftrightarrow \\ &\exists h_k \to 0^+ \quad \exists p_k \xrightarrow{w} p \quad \exists (y_k, y_k^*) \in \text{graph}\,(N_K) \xrightarrow{(n,w)} (y, y^*), \\ &\exists y_k^{*p} \in N_K(y_k + h_k p_k) \quad \text{such that } w - \lim_{k\to\infty} (y_k^{*p} - y_k^*)/h_k = q^*. \end{aligned} \tag{2}$$

Certainly, (1.14), (2.7), (4.42), etc. must also be corrected.

Definition (2) is one "generalized directional derivative" of $N_K(\cdot)$ and it also justifies and states precisely the formal notation $D^2\phi(y^*)p^*$ in Theorem 3.1 of [2] (note that this p^* corresponds to our $-\bar{p}$).

There are two definitions of a derivative of a set-valued map, proposed by [1]. However, they are not suitable for (2). Replacing contingent and Clarke's tangent cones and the norm topology, we adapt the paratingent cone and the weak topology for proposing another definition of a derivative of a set-valued map.

Let X be a Hausdorff topological linear space and $K \subset X$. For any $x \in X$,

$$P_K(x) := \lim_{h\to 0+,\ x'\to x} \sup (K - x')/h \tag{3}$$

is called *paratingent cone* to K at x [3], where lim sup is a Kuratowski limit and

$$\begin{aligned} &v \in P_K(x) \Leftrightarrow \\ &\exists x_\alpha \in K \to x \quad \exists h_\alpha \to 0^+ \quad \exists v_\alpha \to v \quad \forall \alpha,\ x_\alpha + h_\alpha v_\alpha \in K. \end{aligned} \tag{4}$$

Let Y be another Hausdorff topological linear space and F a set-valued map from X to Y. Then, the *paratingent derivative* of F at $(x, y) \in \text{graph}\,(F)$, noted by $PF(x, y)$, is a set-valued map from X to Y, defined by

$$\text{graph}\,(PF(x, y)) = P_{\text{graph}(F)}(x, y). \tag{5}$$

* Received by the editors September 1, 1988; accepted for publication February 6, 1989.
† Nankai Institute of Mathematics, Nankai University, Tianjin, People's Republic of China.

From (3) and (4), it is easy to see that

$$
\begin{aligned}
&v \in PF(x, y)(u) \Leftrightarrow \\
&\exists h_\alpha \to 0^+ \quad \exists u_\alpha \to u \quad \exists (x_\alpha, y_\alpha) \in \text{graph}\,(F), \\
&\exists y_\alpha^v \in F(x_\alpha + h_\alpha u_\alpha) \quad \text{such that } \lim_\alpha (y_\alpha^v - y_\alpha)/h_\alpha = v.
\end{aligned} \tag{6}
$$

Comparing (2) and (6) and letting $X = (V, \sigma(V, V^*))$ and $Y = (V^*, \sigma(V^*, V))$, we obtain that

$$
\begin{aligned}
&D_P N_K(y, y^*) \subset PN_K(y, y^*), \quad \text{i.e.,} \\
&\text{graph}\,(D_P N_K(y, y^*)) \subset \text{graph}\,(PN_K(y, y^*)).
\end{aligned} \tag{7}
$$

However, if $X = (V, \|\cdot\|)$, the inclusions in (7) are inverted.

When $V = V^* = X$ is a Hilbert space, Proposition 7.2.11 of [1] also holds for PN_K or $D_P N_K$, i.e., for instance,

$$q \in D_P N_K(x, p)(u) \Leftrightarrow u \in D_P \pi_K(x + p)(u + q)$$

where $\pi_K : X \to K$ is the projector of best approximation onto K. Furthermore, noting that $\bar{p} \in T_K(\bar{y})$ and $\langle -F(\bar{y}) - E(\bar{u}), \bar{p}\rangle_{V^*, V} = 0$, (1) is quite interesting in this case.

To prove (1), we also need the conception of codifferential for a set-valued map with closed convex graph. According to Definition 4.2.1 of [1], the codifferential $DS(y, x)^*$ of a set-valued map S from Banach space Y to Banach space X with closed convex graph at (y, x) is defined by

$$p \in DS(y, x)^*(q) \Leftrightarrow (p, -q) \in N_{\text{graph}(S)}(y, x). \tag{8}$$

The following proposition is Corollary 4.5.3 of [1].

PROPOSITION 1. *Let X and Y be two Banach spaces, $U : X \to \mathbf{R} \cup \{+\infty\}$ a proper lower semicontinuous, convex function and S a set-valued map from Y to X with closed convex graph. We assume that*

$$0 \in \text{Int}\,(\text{Im}\, S - \text{Dom}\, U).$$

Let $W : Y \to \mathbf{R} \cup \{+\infty\}$ be the marginal function defined by

$$W(y) = \inf_{x \in S(y)} U(x)$$

and let $\bar{x} \in S(\bar{y})$ achieve the minimum of U on $S(\bar{y})$. The subdifferential of the marginal function W is equal to

$$\partial W(\bar{y}) = DS(\bar{y}, \bar{x})^* \, \partial U(\bar{x}).$$

Proof. At first, for simplicity, we assume that $M_n(\cdot, \cdot)$, defined by (4.9), is a continuous single-valued map for the norm topology of $K \times U_{\text{ad}}$ and the weak topology of $N_n B_V$. If K and B_V are strictly convex, this assumption holds. Then, setting

$$p_{nk}^{sw} = M_n(y_n + t_k s, u_n + t_k w),$$

we must replace (4.10) by

$$
\begin{aligned}
&g(y_n + t_k s) + h(u_n + t_k w) - \langle F(y_n + t_k s) + E(u_n + t_k w), p_{nk}^{sw}\rangle_{V^*, V} \\
&\quad > g(y_n) + h(u_n) - \langle F(y_n) + E(u_n), p_{nk}^{sw}\rangle_{V^*, V} - \varepsilon_n t_k (\|s\|_V^2 + \|w\|_U^2)^{1/2} \\
&\quad + \inf_{p \in N_n[(K - y_n - t_k s) \cap B_V]} \langle F(y_n) + E(u_n), p\rangle_{V^*, V} \\
&\quad - \inf_{p \in N_n[(K - y_n) \cap B_V]} \langle F(y_n) + E(u_n), p\rangle_{V^*, V}.
\end{aligned} \tag{9}
$$

Define a set-valued map S_n from V to V by

$$S_n(x) := N_n[(K - y_n - x) \cap B_V], \tag{10}$$

a function $U_n : V \to \mathbf{R}$ by

$$U_n(p) := \langle F(y_n) + E(u_n), p\rangle_{V^*,V}, \tag{11}$$

and a marginal function $W_n : V \to \mathbf{R} \cup \{+\infty\}$ by

$$W_n(x) := \inf_{p \in S_n(x)} U_n(p). \tag{12}$$

Then, (4.15) must be replaced by

$$\begin{aligned} &g^0(y_n; s) + h'(u_n; w) - \langle F'(y_n)s + E'(u_n)w, p_n\rangle_{V^*,V} - W_n'(0; s) \\ &\quad \geqq -\varepsilon_n(\|s\|_V^2 + \|w\|_U^2)^{1/2} \end{aligned} \tag{13}$$

where $p_n = M_n(y_n, u_n)$. Thus, replacing (4.19) and (4.20), we have that

$$\forall s \in K - y_n, \quad g^0(y_n; s) - \langle F'(y_n)^* p_n, s\rangle_{V^*,V} - W_n'(0; s) \geqq -\varepsilon_n \|s\|_V, \tag{14}$$

$$\forall w \in U_{\text{ad}} - u_n, \quad h'(u_n; w) - \langle E'(u_n)^* p_n, w\rangle_{U^*,U} \geqq -\varepsilon_n \|w\|_U. \tag{15}$$

Applying Proposition 1, we have that

$$\partial W_n(0) = DS_n(0, p_b)^*(F(y_n) + E(u_n)). \tag{16}$$

So, from (14) and (16), we obtain that

$$F'(y_n)^* p_n \in N_K(y_n) - DS_n(0, p_n)^*(F(y_n) + E(u_n)) + \partial g(y_n) + \varepsilon_n B_{V^*}. \tag{17}$$

By using the same method as in our paper, we still have (4.29), i.e.,

$$y_n \to \bar{y} \quad \text{in } K \text{ strongly.} \tag{18}$$

We must show that $\{p_n\}$ is bounded and after that, we can assume that

$$p_n \to \bar{p} \quad \text{in } V \text{ weakly,} \tag{19}$$

and deduce (4.34), i.e.,

$$u_n \to \bar{u} \quad \text{in } U_{\text{ad}} \text{ strongly,} \tag{20}$$

provided by extracting a subsequence. In fact, by (14) and (4.31), we have that

$$g^0(y_n; p_n) + \varepsilon_n \|p_n\|_V - W_n'(0, p_n) \geqq \langle F'(y_n)p_n, p_n\rangle_{V^*,V} \geqq C\|p_n\|_V^2. \tag{21}$$

We must show that

$$-W_n'(0; p_n) \leqq O(\|p_n\|_V). \tag{22}$$

Indeed, for $t > 0$ and for any $x \in S_n(tp_n)$, by (10), we have

$$x + tN_n p_n \in N_n(K - y_n)$$

and

$$x + tN_n p_n \in N_n B_V + tN_n p_n \subset N_n(1 + t\|p_n\|_V)B_V$$

and then,

$$\begin{aligned} x + tN_n p_n &\in N_n(K - y_n) \cap N_n(1 + t\|p_n\|_V)B_V \subset N_n(1 + t\|p_n\|_V)\{(K - y_n) \cap B_V\} \\ &= (1 + t\|p_n\|_V)S_n(0). \end{aligned}$$

Hence,

$$S_n(tp_n) \subset (1+t\|p_n\|_V)S_n(0) - tN_n p_n$$

and thus, from (11) and (12), we have that

$$\begin{aligned} W_n'(0, p_n) &= \lim_{t\to 0^+} \frac{\inf_{p\in S_n(tp_n)} U_n(p) - \inf_{p\in S_n(0)} U_n(p)}{t} \\ &\geqq \frac{\inf_{p\in(1+t\|p_n\|_V)S_n(0)} U_n(p) - \inf_{p\in S_n(0)} U_n(p) - tN_n U_n(p_n)}{t} \\ &= -\|p_n\|_V \cdot N_n\| - F(y_n) - E(u_n)\|_{y_n}^K + N_n\langle -F(y_n) - E(u_n), p_n\rangle \\ &\geqq -\|p_n\|_V \cdot N_n\| - F(y_n) - E(u_n)\|_{y_n}^K. \end{aligned}$$

By using (4.22), we deduce (22), and from (21) and (22), it follows that $\{p_n\}$ is bounded.

Now we consider (17). From (18) and (19), we know that $\{F'(y_n)^* p_n\}$ converges weakly. On the other hand, since $\partial g(\cdot)$ is locally bounded, $\{\partial g(y_n)\}$ is relatively weakly compact. Hence, by extracting a subsequence, we can assume that there are two sequences $u_n^* \in N_K(y_n)$ and $v_n^* \in DS_n(0, p_n)^*(F(y_n)+E(u_n))$ such that $w-\lim_{n\to\infty}(u_n^* - v_n^*) = w^*$ exists. We show that for large n,

$$v_n^* \in N_n \cdot N_K\left(y_n + \left(\frac{1}{N_n}\right)p_n\right). \tag{23}$$

In fact, from (8), we have that

$$(v_n^*, F(y_n) - E(u_n)) \in N_{\text{graph}(S_n)}(0, p_n). \tag{24}$$

Set

$$\hat{S}_n(v) := N_n(K - y_n - v) \quad \text{and} \quad B_n(v) \equiv N_n B_V. \tag{25}$$

Then from (10) it follows that

$$S_n(v) = \hat{S}_n(v) \cap B_n(v) \tag{26}$$

and, joining up with (25) and Theorem 4.1.16 of [1],

$$(v_n^*, -F(y_n) - E(u_n)) \in N_{\text{graph}(\hat{S}_n)}(0, p_n) + N_{\text{graph}(B_n)}(0, p_n). \tag{27}$$

But $\{p_n\}$ is bounded, and so for large n

$$(0, p_n) \in \text{Int graph}\,(B_n).$$

Hence,

$$N_{\text{graph}(B_n)}(0, p_n) = \{0\}$$

and thus from (27) we obtain that, for large n,

$$(v_n^*, -F(y_n) - E(u_n)) \in N_{\text{graph}(\hat{S}_n)}(0, p_n). \tag{28}$$

By (25), it means that

$$\forall v \in V \quad \forall w \in N_n(K - y_n - v), \quad \langle v_n^*, v\rangle_{V^*,V} - \langle F(y_n) + E(u_n), w - p_n\rangle_{V^*,V} \leqq 0 \tag{29}$$

and follows that

$$v_n^* = -N_n(F(y_n) + E(u_n)) \tag{30}$$

and

$$(31)\qquad \forall w \in N_n(K - y_n),\quad \langle -(F(y_n) + E(u_n)),\, w - p_n\rangle_{V^*,V} \leqq 0,\quad \text{i.e.,}$$

$$-(F(y_n) + E(u_n)) \in N_K\left(y_n + \frac{1}{N_n} p_n\right),$$

which, indeed, is the definition of p_n. Therefore, from (30) and (31), (23) holds.

Finally, since we also have $u_n^*/N_n \in N_K(y_n)$, by the definition (2), we conclude $w^* \in -D_P N_K(\bar{y}, -F(\bar{y}) - E(\bar{u}))(\bar{p})$, i.e., we complete the proof for the case in which all $M_n(\cdot,\cdot)$ are single-valued.

Now we exclude the singleton assumption of $M_n(\cdot,\cdot)$. Define

$$L_n(y, u) := N_n \| -F(y) - E(u) \|_y^K.$$

Then, by using Proposition 3.2.24 of [1], it is easy to see that $L_n(y, u)$ is locally Lipschitz in a neighborhood of $K \times U_{\text{ad}}$. Hence, (4.15) may be replaced by

$$(32)\qquad g^0(y_n; s) + h'(u_n; w) + L_n^0(y_n, u_n; s, w) \geqq -\varepsilon_n(\|s\|_V^2 + \|w\|_U^2)^{1/2}$$

where $L_n^0(y_n, u_n; s, w)$ is a Clarke directional derivative of L_n, and thus, for any $s \in K - y_n$ and $w \in U_{\text{ad}} - u_n$, there exists $q_n^{sw*} \in \partial L_n(y_n, u_n)$ such that

$$(33)\qquad g^0(y_n; s) + h'(u_n; w) + \langle q_n^{sw*}, (s, w)\rangle_{V^* \times U^*, V \times U} \geqq -\varepsilon_n(\|s\|_V^2 + \|w\|_U^2)^{1/2}.$$

Consider the function $Q : [(K - y_n) \times (U_{\text{ad}} - u_n)] \times \partial L_n(y_n, u_n) \to \mathbf{R}$, defined by

$$(34)\quad Q(s, w; q^*) := g^0(y_n; s) + h'(u_n; w) + \langle q^*, (s, w)\rangle_{V^* \times U^*, V \times U} + \varepsilon_n(\|s\|_V^2 + \|w\|_U^2)^{1/2}.$$

Then, by using the same method as in our paper, we can conclude that there exists $q_n^* \in \partial L_n(y_n, u_n)$ such that

$$\inf_{(s,w)} Q(s, w; q_n^*)$$

$$= \inf_{(s,w)} [g^0(y_n; s) + h'(u_n; w) + \langle q_n^*, (s, w)\rangle_{V^* \times U^*, V \times U} + \varepsilon_n(\|s\|_V^2 + \|w\|_U^2)^{1/2}] \geqq 0,$$

i.e.,

$$(35)\qquad \begin{aligned} &\forall s \in K - y_n \quad \forall w \in U_{\text{ad}} - u_n, \\ &g^0(y_n; s) + h'(u_n; w) + \langle q_n^*, (s, w)\rangle_{V^* \times U^*, V \times U} \geqq -\varepsilon_n(\|s\|_V^2 + \|w\|_U^2)^{1/2}. \end{aligned}$$

We must know what $\partial L_n(y_n, u_n)$ is. We will use a proposition as follows.

PROPOSITION 2. *Let $S_n(\cdot,\cdot)$ be a set-valued map from $V \times V$ to V, defined by*

$$S_n(x, z) := N_n[(K - y_n - z - x) \cap B_V]$$

$$U_n(z, v, p) := \langle F(y_n + z) + E(u_n + v), p\rangle_{V^*,V}$$

and a marginal function

$$W_n(z, v, x) := \inf_{p \in S_n(z,v,x)} U_n(z, v, x).$$

Then W_n is a locally Lipschitz function on a neighborhood of $(K - y_n) \times (U_{\text{ad}} - u_n) \times \{0\}$ and its partial generalized gradient with respect to x at $(z, v, 0)$, $\partial_x W_n(z, v, 0)$, is locally bounded and upper semicontinuous with respect to (z, v) for the norm topology of $(K - y_n) \times (U_{\text{ad}} - u_n)$ and the weak topology of V^*.

The first part of this proposition is also from Proposition 3.2.24 of [1] and its second part is well known.

Now, suppose that for $t_k \to 0^+$, $z_k \in V \to 0$, and $v_k \in U \to 0$,

$$L_n^0(y_n, u_n; s, w) = \lim_{k\to\infty} \{L_n(y_n + z_k + t_k s, u_n + v_k + t_k w) - L_n(y_n + z_k, u_n + v_k)\}/t_k.$$

Since

$$\begin{aligned}
&L_n(y_n + z_k + t_k s, u_n + v_k + t_k w) - L_n(y_n + z_k, u_n + v_k) \\
&\quad \leqq \langle -F(y_n + z_k + t_k s) - E(u_n + v_k + t_k w), \hat{p}_{nk}^{sw} \rangle_{V^*,V} \\
&\qquad - \langle -F(y_n + z_k) - E(u_n + v_k), \hat{p}_{nk}^{sw} \rangle_{V^*,V} \\
&\qquad - \inf_{p \in N_n\{(K - y_n - z_k - t_k s) \cap B_V\}} \langle F(y_n + z_k) + E(u_n + v_k), p \rangle_{V^*,V} \\
&\qquad + \inf_{p \in N_n\{(K - y_n - z_k) \cap B_V\}} \langle F(y_n + z_k) + E(u_n + v_k), p \rangle_{V^*,V} \\
&\quad = \langle -F(y_n + z_k + t_k s) - E(u_n + v_k + t_k w), \hat{p}_{nk}^{sw} \rangle_{V^*,V} \\
&\qquad - \langle -F(y_n + z_k) - E(u_n + v_k), \hat{p}_{nk}^{sw} \rangle_{V^*,V} - W_n(z_k, v_k, t_k s) + W_n(z_k, v_k, 0)
\end{aligned}$$

where $\hat{p}_{nk}^{sw} \in M_n(y_n + z_k + t_k s, u_n + v_k + t_k w)$ and $W_n(\cdot, \cdot, \cdot)$ is defined as in Proposition 2. Obviously,

$$\begin{aligned}
&\lim_{k\to\infty} \{\langle -F(y_n + z_k + t_k s) - E(u_n + v_k + t_k w), \hat{p}_{nk}^{sw} \rangle_{V^*,V} \\
&\qquad - \langle -F(y_n + z_k) - E(u_n + v_k), \hat{p}_{nk}^{sw} \rangle_{V^*,V}\}/t_k = \langle -F'(y_n)s - E'(u_n)w, p_n^{sw} \rangle_{V^*,V}
\end{aligned}$$

where $p_n^{sw} \in M_n(y_n, u_n)$. On the other hand, by the Mean-Value Theorem, there exist $\theta_k \in (0, 1)$ and $r_{nk}^* \in \partial_x W_n(z_k, v_k, \theta_k t_k s)$ such that

$$W_n(z_k, v_k, t_k s) - W_n(z_k, v_k, 0) = \langle r_{nk}^*, t_k s \rangle_{V^*,V}.$$

Using Proposition 2 and extracting a subsequence, we can assume that

$$w - \lim_{k\to\infty} r_{nk}^* = r_n^* \in \partial_x W_n(0, 0, 0) = \partial W_n(0)$$

where $W_n(\cdot)$ is defined as (12) and

$$\lim_{k\to\infty} \{W_n(z_k, v_k, t_k s) - W_n(z_k, v_k, 0)\}/t_k = \langle -r_n^*, s \rangle_{V^*,V} \leqq -W_n'(0; s).$$

Thus, we obtain that

$$L_n^0(y_n, u_n; s, w) \leqq \langle -F'(y_n)s - E'(u_n)w, p_n^{ws} \rangle_{V^*,V} - W_n'(0; s)$$

with $p_n^{ws} \in M_n(y_n, u_n)$, and then, for any s and w,

$$L_n^0(y_n, u_n; s, w) \leqq \sup_{p \in M_n(y_n, u_n)} \{\langle -F'(y_n)^* p, s \rangle_{V^*,V} + \langle -E'(u_n)^* p, w \rangle_{U^*,U}\} - W_n'(0; s).$$

It follows that

$$\partial L_n(y_n, u_n) \subset \bigcup_{p \in M_n(y_n, u_n)} (-F'(y_n)^* p, -E'(u_n)^* p) - (\partial W_n(0), 0).$$

Therefore, for any $q_n^* \in \partial L_n(y_n, u_n)$, there exists a $p_n \in M_n(y_n, u_n)$ such that

$$q_n^* \in (-F'(y_n)^* p_n, -E'(u_n)^* p_n) - (\partial W_n(0), 0).$$

Joining with (35), it deduces that (14) and (15) hold again. Furthermore, by using Proposition 1, (16) holds for every $p_n \in M_n(y_n, u_n)$. Hence, (17) holds also. Thus, it reduces to the first case.

REFERENCES

[1] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.

[2] V. BARBU, *Optimal Control of Variational Inequalities*, Pitman, Boston, 1984.

[3] SHUZHONG SHI, *Théorème de Choquet et Analyse non régulière*, C.R. Acad. Sci. Paris Ser. I Math., 305 (1987), pp. 41-44. *Choquet Theorem and nonsmooth analysis*, J. Math. Pure. Appl., 67 (1988), pp. 411-432.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 250-251, January 1990

013

ERRATUM AND ADDENDUM: Positive Semidefinite Matrices: Characterization via Conical Hulls and Least-Squares Solution of a Matrix Equation*

J. C. ALLWRIGHT† AND K. G. WOODGATE‡

Contrary to Theorem 3.1, the optimization problem

$$\text{(P1)} \qquad \min_{A \in S_{\geqq}^n} \|F - AG\|_F$$

does not necessarily have a solution when $\text{rank}\,[G] < n$, where $F, G \in R^{n\times m}$ and $S_{\geqq}^n = \{A \in R^{n\times n}: A' = A \geqq 0\}$. For example, consider Woodgate's counterexample, which initiated this note:

$$F = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \qquad G = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

for which $\inf_{A \in S_{\geqq}^2} \|F - AG\|_F = 1$ but is not achieved by any $A \in S_{\geqq}^2$.

Theorem 3.1 should actually have been as follows.

THEOREM 3.1′. *The minimum in* (P1) *exists when* $\text{rank}\,[G] = n$.

The *error* in the proof of Theorem 3.1 for the case $\text{rank}\,[G] < n$ is caused by the incorrect statement in the penultimate paragraph of page 546 that the set $\text{vec}\,(S_{BC0})$ is closed, where S_{BC0}, of (3.31), is defined by

$$S_{BC0} = \left\{ \begin{bmatrix} B & C \\ C' & 0 \end{bmatrix} : \begin{bmatrix} B & C \\ C' & D \end{bmatrix} \in S_{\geqq}^n \right\}.$$

This set, and hence $\text{vec}\,(S_{BC0})$, is not closed because, for any nonzero C, the matrix $\begin{bmatrix} 0 & C \\ C' & 0 \end{bmatrix}$ is a limit point of S_{BC0} which is not in S_{BC0}.

Suppose now that $\text{rank}\,[G] < n$.

If (P1) actually has a minimum, then the results of § 3 apply. If (P1) does not have a minimum, then it is easy to check that those results, and their proofs, still apply when all occurrences of the nonexistent minimal value $\|F - \hat{A}G\|_F$ are replaced by $\inf\{\|F - AG\|_F: A \in S_{\geqq}^n\}$. Then the methodology of § 3 for finding an $\tilde{A} \in S_{\geqq}^n$ which solves (P1) to prespecified accuracy actually yields an $\tilde{A}$ such that $\|F - \tilde{A}G\|_F$ approximates $\inf\{\|F - AG\|_F: A \in S_{\geqq}^n\}$ to prespecified accuracy—which, from the practical point of view, is just as good.

It is possible to say more for the case when $\text{rank}\,[G] < n$.

From (3.37), for $A \in S_{\geqq}^n$,

$$\text{(A1)} \qquad \|F - AG\|_F^2 = \|F_1 - BG_1\|_F^2 + \|F_2 - C'G_1\|_F^2$$

when A is partitioned as

$$A = Q' \begin{bmatrix} B & C \\ C' & D \end{bmatrix} Q.$$

* Received by the editors July 18, 1988; accepted for publication (in revised form) May 26, 1989. SIAM J. Control Optim., 26 (1988), pp. 537-556.

† Department of Electrical Engineering, Imperial College of Science and Technology, London SW7 2BT, United Kingdom.

‡ University of Twente, Post Office Box 217, 7500 Enschede, the Netherlands.

Since rank $[G_1] = q$ and $B \in S_{\geqq}^q$, it follows from Theorem 3.1′ above that there is a $\hat{B}$ which minimizes $\|F_1 - BG_1\|_F^2$ with respect to $B \in S_{\geqq}^q$. If $\tilde{C} = (F_2 G_1^\dagger)'$ and $R[\tilde{C}] \subset R[\hat{B}]$, then

$$A^* = Q' \begin{bmatrix} \hat{B} & \tilde{C} \\ \tilde{C}' & D \end{bmatrix} Q$$

is optimal for (P1) for any $D \in S_{\geqq}^{n-q}$ such that $D \geqq \tilde{C}' \hat{B}^\dagger \tilde{C}$. This occurs because it follows from (3.36) that $A^* \in S_{\geqq}^n$ and from (A1) above that $\hat{B}$ and $\tilde{C}$ minimize the right-hand side of (A1) with respect to $B \in S_{\geqq}^q$ and $C \in R^{q \times (n-q)}$. Hence (P1) certainly has a minimum when $R[\tilde{C}] \subset R[\hat{B}]$. In fact the following holds.

THEOREM A1. *If* rank $[G] < n$, *then* (P1) *has a minimum if and only if* $R[\tilde{C}] \subset R[\hat{B}]$.

Proof. In view of the above, it just remains to be shown that there is no minimum when $R[\tilde{C}] \not\subset R[\hat{B}]$.

Suppose that $R[\tilde{C}] \not\subset R[\hat{B}]$ and that there is a minimum, say at $A^0 \in S_{\geqq}^n$. Partition A^0 as

$$A^0 = Q' \begin{bmatrix} B^0 & C^0 \\ C^{0\prime} & D^0 \end{bmatrix} Q,$$

where, from (3.36), $B^0 \in S_{\geqq}^q$ and $R[C^0] \subset R[B^0]$.

If $C^0 = \tilde{C}$ then $B^0 \neq \hat{B}$. For suppose $B^0 = \hat{B}$. Then $R[C^0] \subset R[B^0] = R[\hat{B}]$, so that, because the case $C^0 = \tilde{C}$ is being considered, $R[\tilde{C}] \subset R[\hat{B}]$, which contradicts the initial assumption that $R[\tilde{C}] \not\subset R[\hat{B}]$. Since rank $[G_1] = q$, Theorem 3.2 reveals that $\hat{B}$ is the unique global minimizer of $\|F_1 - BG_1\|_F^2$ with respect to $B \in S_{\geqq}^q$ so, since $B^0 \in S_{\geqq}^q$ and $B^0 \neq \hat{B}$, $\|F_1 - B^0 G_1\|_F^2 > \|F_1 - \hat{B} G_1\|_F^2$.

On the other hand, if $C^0 \neq \tilde{C}$ then $\|F_2 - C^{0\prime} G_1\|_F^2 > \|F_2 - \tilde{C}' G_1\|_F^2$, because, since rank $[G_1] = q$, $\tilde{C}$ is the unique minimizer of $\|F_2 - C' G_1\|_F^2$ with respect to $C \in R^{q \times (n-q)}$. Also $\|F_1 - B^0 G_1\|_F^2 \geqq \|F_1 - \hat{B} G_1\|_F^2$ owing to the optimality of $\hat{B}$.

Hence, whether $C^0 = \tilde{C}$ or $C^0 \neq \tilde{C}$,

$$\begin{aligned} \|F - A^0 G\|_F^2 &= \|F_1 - B^0 G_1\|_F^2 + \|F_2 - C^{0\prime} G_1\|_F^2 \\ &> \|F_1 - \hat{B} G_1\|_F^2 + \|F_2 - \tilde{C}' G_1\|_F^2 \\ &= \inf \{\|F - AG\|_F^2 : A \in S_{\geqq}^n\} \end{aligned}$$

where the first equality is from (A1) and the last equality is from (3.12) with $\|F - \hat{A} G\|_F^2$ replaced by $\inf \{\|F - AG\|_F^2 : A \in S_{\geqq}^n\}$, as mentioned earlier in this note. This contradicts the optimality of A^0 and therefore contradicts the existence of a minimum for (P1) when $R[\tilde{C}] \not\subset R[\hat{B}]$, which completes the proof. □

A final point is that it is possible to slightly modify the perturbations made in Theorem 3.5 in order to cause $\tilde{B}$ to be positive definite, as it is only necessary that they cause $\tilde{B}$ to be positive semidefinite with $R[\tilde{C}] \subset R[\tilde{B}]$.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 253-264, March 1990

OPTIMAL CONTROL FOR AN INFINITE-DIMENSIONAL PERIODIC PROBLEM UNDER WHITE NOISE PERTURBATIONS*

CONSTANTIN TUDOR†

Abstract. In this paper a result of the type of "law of large numbers" is obtained for the infinite-dimensional version of the linear quadratic cost problem in the periodic case if the deterministic optimal feedback law is used in the presence of white noise perturbations.

Key words. optimal control, periodic problem, law of large numbers

AMS(MOS) subject classifications. 49B, 93E

1. Introduction. In this paper we are concerned with a stochastic linear infinite-dimensional control system of Ito type with periodic coefficients.

We associate an optimization problem, which is natural for the periodic case, with the cost $\int_0^\theta y(t)\,dt$, where θ is the period and y is a process depending quadratically on the control. Now it is clear in the usual deterministic situation that

$$\int_0^\theta y(t)\,dt = \lim_{n\to\infty} \frac{1}{n}\int_0^{n\theta} y(t)\,dt$$

and this suggests that in the stochastic case we use results of the type of the "law of large numbers" by considering cost functions of the form $\overline{\lim}_{n\to\infty}(1/n)\int_0^{n\theta} y(t)\,dt$.

In the finite-dimensional case such results of the "law of large numbers" type have been considered by Mandl [15] for stationary systems and by Halanay, Tudor, and Morozan [11] for almost periodic systems. The periodic infinite-dimensional deterministic case with a cost of the form $\int_0^\theta y(t)\,dt$ has been considered by Da Prato [5]; the periodic infinite-dimensional stochastic case with a cost of the form $E\int_0^\theta y(t)\,dt$ has been considered by Da Prato and Ichikawa [7], [8]; Da Prato and Ichikawa [6] have also considered the case of almost periodic forcing terms, with the cost $\lim_{T\to\infty}(1/T)E\int_0^T y(t)\,dt$. The optimal control has been obtained by Da Prato in [5] as an affine function of the state with the aid of the periodic solution of a Riccati equation and it has been seen that this control is still optimal in the corresponding stochastic setting with expected cost.

The main contribution of the present paper is that it shows that the same control is optimal almost surely with respect to the cost $\overline{\lim}_{n\to\infty}(1/n)\int_0^{n\theta} y(t)\,dt$.

2. Notation and hypotheses. Let M, K, U be real separable Hilbert spaces and let $\theta > 0$. $L(K, H)$ (shortly $L(H)$ if $H = K$) denote the Banach space of all bounded linear operators from K to H. $\Sigma(H) = \{\Pi \in L(H);\ \Pi = \Pi^*\}$, $\Sigma^+(H) = \{\Pi \in \Sigma(H);\ \Pi \geqq 0\}$, where Π^* represents the adjoint of Π. By Ex we denote the expectation of the random variable x, by $\operatorname{Tr}\Pi$ the trace of the operator Π and if X is a topological space we will denote by $\mathcal{B}_X$ the σ-algebra of Borel sets in X. Let $(\Omega, \mathcal{F}, P, (\mathcal{F}_t)_{t\in R})$ be a filtered probability space and let $\{w(t)\}_{t\in R}$ be a K-valued $\mathcal{F}_t$-Wiener process with W as covariance operator ([16]).

DEFINITION. $U(\cdot,\cdot):\{(t, s);\ 0 \leqq s \leqq t\} \to L(H)$ is an evolution operator if

(2.1) $U(t, t) = I$, the idenity operator, for all t,

* Received by the editors December 4, 1987; accepted for publication (in revised form) April 24, 1989.
† Faculty of Mathematics, University of Bucharest, Street Academiei 14, 70109 Bucharest, Romania.

(2.2) $U(t, r)U(r, s) = U(t, s), 0 \leqq s \leqq r \leqq t,$

(2.3) $U(t, s)$ is strongly continuous in s on $[0, t]$ and strongly continuous in t on $[s, \infty)$,

(2.4) For every $T > 0$ there is a constant C_T such that

$$\| U(t, s) \|_{L(H)} \leqq C_T, 0 \leqq s \leqq t \leqq T.$$

It should be noted that (2.4) does not follow from (2.1)-(2.3) as is sometimes supposed (see [10] for a counterexample).

DEFINITION. A strong evolution operator is an evolution operator for which there exists a closed, linear, densely defined operator $A(t)$, $t \geqq 0$, with the domain $D(A(t))$, such that

(2.5) $$U(t, s) : D(A(s)) \to D(A(t)) \quad \text{for } t > s,$$

(2.6) $$\frac{\partial}{\partial t} U(t, s)h = A(t)U(t, s)h \quad \text{for } h \in D(A(s)), \quad t > s.$$

We will assume the following hypotheses (such as Da Prato [5]):

(2.7) (a) For every $t \in R$, $A(t)$ is a closed, linear, densely defined operator and the map $t \to A(t)$ is θ-periodic;
(b) $A(t)$ generates a strong evolution operator $\{U(t, s)\}$;
(c) There exists the Yosida approximation $A_n(t) = n^2[n - A(t)]^{-1} - nI$ for n sufficiently large. Moreover, if $g \in L^2([0, \theta], H)$ and $z(t) = U(t, 0)x + \int_0^t U(t, s)g(s)\, ds$, z_n is the strict solution of $z'(t) = A_n(t)z_n(t) + g(t)$, $z_n(0) = x$, then $\sup_{t \leqq \theta} |z_n(t) - z(t)| \to 0$.

Remark 2.1. From (2.7)(a)-(c) we have that $U(t + \theta, s + \theta) = U(t, s)$ for all $t > s$.

Remark 2.2. Conditions (2.7)(b), (c) are fulfilled if the usual hypotheses of Tanabe and Kato-Tanabe are satisfied [19]:

(2.8) (a) $B : R \to L(U, H)$, $G : R \to L(K, H)$ are θ-periodic and strongly continuous,
(b) $f : R \to H$ is θ-periodic and $f \in L^2([0, \theta], H)$,
(c) $M : R \to \Sigma^+(H)$, $N : R \to \Sigma^+(U)$ are θ-periodic, strongly continuous, and $N(t) \geqq \gamma I$, $\gamma > 0$, for all t.

(2.9) 1 belongs to the resolvent set of $U(\theta, 0)$.

(2.10) *Stabilizability.* There exist a θ-periodic function $C : R \to L(H, U)$ strongly continuous and $\alpha, \beta > 0$ such that $\| U_{A-BC}(t, s) \|_{L(H)} \leqq \alpha \exp\{-\beta(t - s)\}$ for all $s \leqq t$, where U_{A-BC} is the evolution operator relative to $A - BC$.

(2.11) *Detectability.* There exists a θ-periodic function $C_1 : R \to L(H)$ strongly continuous and $\alpha_1, \beta_1 > 0$ such that $\| U_{A-C_1M^{1/2}}(t, s) \|_{L(H)} \leqq \alpha_1 \exp\{-\beta_1(t - s)\}$ for all $s \leqq t$, where $U_{A-C_1M^{1/2}}$ is the evolution operator relative to $A - C_1M^{1/2}$.

3. A class of random processes in Hilbert spaces. Let ξ be a H-valued $\mathcal{F}_a$-measurable random element $(a \in R)$ and let $\{V(t, s)\}$ be an evolution operator that is θ-periodic, i.e., $V(t + \theta, s + \theta) = V(t, s)$ for all $s \leqq t$. Let $g : R \to H$ be continuous, θ-periodic, and let $G : R \to L(K, H)$ be strongly continuous and θ-periodic.

PROPOSITION 3.1. *Under the above assumptions the process*

$$x(t,\xi)=V(t,a)\xi+\int_a^t V(t,s)g(s)\,ds+\int_a^t V(t,s)G(s)\,dw(s),\qquad t\geqq a \tag{3.1}$$

is a Markov process with the transition function $P(s,h,t,A)$ θ*-periodic, i.e.,* $P(s+\theta,h,t+\theta,A)=P(s,h,t,A)$ *for all* $a\leqq s\leqq t$, $h\in H$, $A\in\mathcal{B}_H$. *Moreover, if* V *is exponentially stable, i.e., if there is* α_2, $\beta_2>0$ *such that*

$$\|V(t,s)\|_{L(H)}\leqq\alpha_2\exp\{-\beta_2(t-s)\}\quad \textit{for all } s\leqq t, \tag{3.2}$$

and $E(|\xi|^4)<\infty$, *then* $\sup_{t\geqq a}E(|x(t,\xi)|^4)<\infty$.

Proof. The Markov property of $x(t,\xi)$ is shown in [1]. Let $x(t,s,\xi)_{t\geqq s}$ be the process defined by

$$x(t,s,\xi)=V(t,s)\xi+\int_s^t V(t,u)g(u)\,du+\int_s^t V(t,u)G(u)\,dw(u). \tag{3.3}$$

Then $P(s,h,t,A)=P(x(t,s,h)\in A)$ and the θ-periodicity of the transition function follows at once from the θ-periodicity of V, g, G and the stationarity of w.

Assume now that V is stable and $E(|\xi|^4)<\infty$. We have

$$E(|V(t,a)\xi|^4)\leqq\alpha_2^4E(|\xi|^4),$$

$$\left|\int_0^t V(t,s)g(s)\,ds\right|\leqq\frac{\alpha_2}{\beta_2}\sup_s|g(s)|.$$

Recall that if y is a Gaussian random element with mean zero and covariance Π, then

$$E(|y|^{2n})\leqq(2n-1)!!(\operatorname{Tr}\Pi)^n \tag{3.4}$$

for any integer n, where $(2n-1)!!=(2n-1)(2n-3)\cdots 5.3.1$.

Since the stochastic integral $\int_0^t V(t,s)G(s)\,dw(s)$ is Gaussian with mean zero and covariance

$$\int_0^t V(t,s)G(s)WG^*(s)V^*(t,s)\,ds,$$

then by using (3.4) we obtain

$$\begin{aligned}
&E\left(\left|\int_0^t V(t,s)G(s)\,dw(s)\right|^4\right)\\
&\quad\leqq 3\left\{\int_0^t\operatorname{Tr}[V(t,s)G(s)WG^*(s)V^*(t,s)]\,ds\right\}^2\\
&\quad\leqq 3(\operatorname{Tr}W)^2\sup_s\|G(s)\|^4_{L(K,H)}\left[\int_0^t\|V(t,s)\|^2_{L(H)}\,ds\right]^2\\
&\quad\leqq\frac{3\alpha_2^4(\operatorname{Tr}W)^2}{4\beta_2^2}\sup_s\|G(s)\|^4_{L(K,H)}.
\end{aligned}$$

Now it remains to apply the inequality $|x+y+z|^4\leqq 27(|x|^4+|y|^4+|z|^4)$, x, y, $z\in H$.

PROPOSITION 3.2. *Let* $\{x_n(t)\}_{t\geqq -n}$ *be the process defined by* (3.1) *for* $a=-n$ *and* $\xi=0$. *Then we have the following*:

(a) *For every* t, $x_n(t)$ *converges in* $L^2(\Omega,\mathcal{F},P)$ *to* $x(t)$ *and the process* $\{x(t)\}_{t\in R}$ *is measurable,* $\mathcal{F}_t$*-adapted, and* $t\to x(t):R\to L^2(\Omega,\mathcal{F},P)$ *is continuous.*

(b) *The following inequality holds*:

$$\sup_t E(|x(t)|^4) < \infty. \tag{3.5}$$

(c) $\{x(t)\}$ *is a* θ*-periodic Markov process and for* $t \geqq s$

$$x(t) = V(t, s)x(s) + \int_s^t V(t, r)g(r)\, dr + \int_s^t V(t, r)G(r)\, dw(r) \quad P\text{-a.s.} \tag{3.6}$$

We will denote the process $x(t)$ *by*

$$x(t) = \int_{-\infty}^t V(t, s)g(s)\, ds + \int_{-\infty}^t V(t, s)G(s)\, dw(s). \tag{3.7}$$

We proceed as in [17].

Proof. (a) Recall that

$$\sup_{t \geqq -n} E(|x_n(t)|^4) \leqq \gamma_1 < \infty \quad \text{(see Proposition 3.1).} \tag{3.8}$$

If $t \geqq -n > -m$ we have

$$\begin{aligned} E(|x_m(t) - x_n(t)|^2) &= E(|V(t, -n)x_m(-n)|^2) \\ &\leqq \alpha_2^2 \exp\{-2\beta_2(t+n)\} E(|x_m(-n)|^2) \\ &\leqq \gamma_2 \exp(-2\beta_2 n) \to 0 \end{aligned}$$

as $n \to \infty$, uniformly on every interval $[\alpha, \infty)$, $\alpha \in R$. Then the process $x(t) = \lim_{n\to\infty} x_n(t)$ (L^2-limit) is measurable, adapted and $t \to x(t) : R \to L^2(\Omega, \mathcal{F}, P)$ is continuous.

(b) Inequality (3.5) follows from (3.8).

(c) If $t > s \geqq -n$, then we have

$$x_n(t) = V(t, s)x_n(s) + \int_s^t V(t, r)g(r)\, dr + \int_s^t V(t, r)G(r)\, dw(r)$$

where from letting $n \to \infty$ we obtain (3.6).

The Markov property follows from (3.6) as in [1]. The transition function is given as in Proposition 3.1 and is θ-periodic. Next, for every $t \in R$, $h \in H$, we have

$$\lim_{n\to\infty} E(\exp(i\langle h, x_n(t)\rangle)) = E(\exp(i\langle h, x(t)\rangle)),$$

$$\lim_{n\to\infty} E(\exp(i\langle h, x_{n+\theta}(t+\theta)\rangle)) = E(\exp(i\langle h, x(t+\theta)\rangle)),$$

$$\begin{aligned} E(\exp(i\langle h, x_{n+\theta}(t+\theta)\rangle)) &= \int \exp(i\langle h, z\rangle) P(n+\theta, 0, t+\theta, dz) \\ &= \int \exp(i\langle h, z\rangle) P(n, 0, t, dz) \\ &= E(\exp(i\langle h, x_n(t)\rangle)). \end{aligned}$$

Therefore

$$E(\exp(i\langle h, x(t)\rangle)) = E(\exp(i\langle h, x(t+\theta)\rangle))$$

for all $h \in H$, so that $x(t)$ and $x(t+\theta)$ have the same distribution for every t. This fact, together with the Markov property, implies that $\{x(t)\}$ is θ-periodic.

LEMMA 3.3. *Let* $\{h(t)\}_{t\geqq 0}$ *be an H-valued process such that*

$$\sup_t E(|h(t)|^2) < \infty$$

and let $\{R(t)\}_{t\geqq 0}$ *be an* $L(K, H)$*-valued progressively measurable process such that*

$$\sup_t E(\|R(t)\|^4_{L(K,H)}) < \infty.$$

Then P-almost surely we have

(a) $\lim_{n\to\infty} (1/n)h(n) = 0$.

(b) $\lim_{n\to\infty} (1/n)\int_0^n R(t)\, dw(t) = 0$.

Proof. (a) By Chebyshev's inequality for $\varepsilon > 0$ we have

$$P\left(\frac{1}{n}|h(n)| > \varepsilon\right) \leqq \frac{1}{n^2\varepsilon^2} E(|h(n)|^2) \leqq \frac{\gamma_3}{n^2}.$$

Since $\sum_n 1/n^2 < \infty$, we may apply the Borel-Cantelli Lemma.

(b) By using the inequality (see [13])

$$E\left(\left|\int_0^t \varphi(s)\, dw(s)\right|^4\right) \leqq \gamma_4 t \int_0^t E(\|\varphi(s)\|^4_{L(K,H)})\, ds$$

and the Markov inequality, we obtain

$$\begin{aligned} P\left(\frac{1}{n}\left|\int_0^n R(t)\, dw(t)\right| > \varepsilon\right) &\leqq \frac{1}{n^4\varepsilon^4} E\left(\left|\int_0^n R(t)\, dw(t)\right|^4\right) \\ &\leqq \frac{\gamma_5}{n^3}\int_0^n E(\|R(t)\|^4_{L(K,H)})\, dt \\ &\leqq \frac{\gamma_6}{n^2} \end{aligned}$$

and we use the Borel-Cantelli Lemma again.

4. The deterministic optimization problem. We consider the following Riccati equation:

$$Q' + A^*Q + QA - QBN^{-1}B^*Q + M = 0. \tag{4.1}$$

DEFINITION [5]. A strongly continuous function $Q:[0, \theta] \to \Sigma^+(H)$ is a θ-periodic solution of (4.1) if there is $S \in \Sigma^+(H)$ such that $Q(0) = Q(\theta) = S$ and for every $h \in H$

$$\begin{aligned} Q(t)h = U^*(\theta, t)SU(\theta, t)h \\ -\int_t^\theta U^*(s, t)[Q(s)B(s)N^{-1}(s)B^*(s)Q(s) - M(s)]U(s, t)h\, ds. \end{aligned} \tag{4.2}$$

THEOREM 4.1[5]. (a) *Assume* (2.7), (2.8), (2.10). *Then there exists a* θ*-periodic solution* Q *of* (4.1).

(b) *Assume* (2.7)-(2.10) *and*

(4.3) 1 *belongs to the resolvent set of the evolution operators* $U_{L(Q)}$, $U_{L^*(Q)}$ *generated by* $L(Q) = A - BN^{-1}B^*Q$, $L^*(Q) = A^* - QBN^{-1}B^*$.

Consider the following control problem. Minimize the cost functional

$$\bar{J}(u) = \int_0^\theta [\langle M(t)y(t), y(t)\rangle + \langle N(t)u(t), u(t)\rangle]\, dt \tag{4.4}$$

over all $u \in L^2([0, \theta], H)$ *subject to*

$$y'(t) = A(t)y(t) + B(t)u(t) + f(t), \qquad y(0) = y(\theta). \tag{4.5}$$

Then the optimal control is given by

$$\tilde{u} = -N^{-1}B^*(Qy + r) \tag{4.6}$$

and the optimal cost by

$$\bar{J}(\tilde{u}) = \int_0^\theta [2\langle f(s), r(s)\rangle - \langle N^{-1}(s)B^*(s)r(s), B^*(s)r(s)\rangle]\, ds \tag{4.7}$$

where r is the unique solution of

$$r' + (A^* - QBN^{-1}B^*)r + Qf = 0, \qquad r(0) = r(\theta) \tag{4.8}$$

and y is the unique solution of the equation

$$y' = (A - BN^{-1}B^*Q)y - BN^{-1}B^*r + f, \qquad y(0) = y(\theta). \tag{4.9}$$

THEOREM 4.2. *Assume* (2.7), (2.8) *and either* (2.10), (2.11) *or*

$$M(t) \geqq \alpha I, \qquad \alpha > 0, \tag{4.10}$$

for all t, and for each $s \in R$, $x \in H$ *there is u strongly measurable such that* $\int_s^\infty [\langle M(t)y(t), y(t)\rangle + \langle N(t)u(t), u(t)\rangle]\, dt < \infty$. *Then there exists a unique θ-periodic solution Q of* (4.1) *and the evolution operator* $U_{L(Q)}$ *is exponentially stable.*

Proof. Under (2.7), (2.8), (2.10), and (2.11) the first part of the theorem is proved in [5]. From Lemma 3.5 of [5] it follows that there is $C > 0$ such that for all s

$$\int_s^\infty |U_{L(Q)}(t, s)x|^2\, dt \leqq C|x|^2 \quad \text{for every } x \in H.$$

Therefore by a result due to Datko (see [9] or [14]) $U_{L(Q)}$ is exponentially stable.

Under (2.7), (2.8), and (4.10) the result is proved in Theorem 4.9 of [10].

Remark 4.1. Hypothesis (4.3) in Theorem 4.1 can be replaced by the following:

$$U_{L(Q)},\ U_{L^*(Q)} \text{ are exponentially stable.} \tag{4.11}$$

Indeed (4.3) has been used only to obtain a θ-periodic solution of (4.8) and (4.9). Under (4.11) $r(t) = \int_t^\infty U_{L^*(Q)}(s, t)Q(s)f(s)\, ds$ and $y(t) = \int_{-\infty}^t U_{L(Q)}(t, s) \cdot [f(s) - B(s)N^{-1}(s)B^*(s)r(s)]\, ds$ are the required solutions of (4.8) and (4.9).

Remark 4.2 [3, Lemma 1.2, Thm. 2.4]. Suppose (2.7), (2.8), (2.10) are satisfied. Moreover, assume that $s \to \langle U(t, s)A(s)x, y\rangle$ is integrable for all $y \in H$ and $x \in \mathcal{D} = \bigcap_{0 \leqq t \leqq \theta} D(A(t))$.

If Q is a θ-periodic solution of (4.1), then Q satisfies the following inner product Riccati equation:

$$\frac{d}{dt}\langle Qh, h\rangle + \langle [A^*Q + QA + M - QBN^{-1}B^*Q]h, h\rangle = 0 \quad \text{for all } h \in \mathcal{D}. \tag{4.12}$$

We remark that a sufficient condition for $s \to \langle U(t, s)A(s)x, y\rangle$ to be integrable is that it be measurable and $\sup_{t \leqq \theta} |A(t)x| < \infty$ for each $x \in \mathcal{D}$, and this condition is usually satisfied in the applications.

Remark 4.3. For affine and θ-periodic controls u for which the linear part is stabilizing, existence of periodic dynamics is ensured and we have that $\bar{J}(u)$ given by (4.4) satisfies

$$\bar{J}(u) = \lim_{n \to \infty} \frac{1}{n} \int_0^{n\theta} [\langle M(t)y(t), y(t)\rangle + \langle N(t)u(t), u(t)\rangle]\, dt.$$

This is the reason we consider the cost functional (5.5) in the periodic stochastic case below.

5. Main result. In this section we will consider a stochastic control problem with the cost calculated pathwise.

We consider the controlled stochastic evolution equation

$$dx(t) = [A(t)x(t) + B(t)u(t) + f(t)]\,dt + G(t)\,dw(t). \tag{5.1}$$

DEFINITION. A measurable, adapted process x with values in H is a mild solution of (5.1) on R if for every $t \geqq s$

$$x(t) = U(t,s)x(s) + \int_s^t U(t,r)[B(r)u(r) + f(r)]\,dr + \int_s^t U(t,r)G(r)\,dw(r) \quad P\text{-a.s.} \tag{5.2}$$

We denote by U_{ad} the space of all measurable, adapted, and θ-periodic processes $\{u(t)\}_{t\in R}$ in U with

$$\sup_t E(|u(t)|^4) < \infty \tag{5.3}$$

for which (5.1) has a unique (up to a modification) θ-periodic solution with

$$\sup_t E(|x(t)|^4) < \infty. \tag{5.4}$$

The statement of the control problem is the following. Find $u \in U_{\mathrm{ad}}$ that minimizes over U_{ad} (in the sense of the almost surely inequality) the random cost functional

$$J(u) = \overline{\lim_{n\to\infty}} \frac{1}{n} \int_0^{n\theta} [\langle M(t)x(t), x(t)\rangle + \langle N(t)u(t), u(t)\rangle]\,dt. \tag{5.5}$$

It is not clear if there exist any admissible controls. In this sense we have the following results.

PROPOSITION 5.1. *Suppose that $\{y(t)\}_{t\geqq 0}$ is a real measurable and θ-periodic process with $\sup_t E(|y(t)|) < \infty$. Then $(1/n)\int_0^{n\theta} y(s)\,ds$ converges P-almost surely to a finite random variable.*

PROPOSITION 5.2. *Suppose* (2.10) *is satisfied. Then the feedback control $u(t) = -C(t)x(t) + v(t)$, where $v: R \to U$ is continuous and θ-periodic, is in U_{ad}, and in particular $J(u)$ is finite P-almost surely. Moreover, if $\{x_1(t)\}_{t\geqq 0}$ is any solution of* (5.1) *on R_+ with $E(|x_1(0)|^2) < \infty$ corresponding to the affine control $u_1(t) = -C(t)x_1(t) + v(t)$, then the cost $\overline{\lim}_{n\to\infty}(1/n)\int_0^{n\theta}[\langle M(t)x_1(t), x_1(t)\rangle + \langle N(t)u_1(t), u_1(t)\rangle]\,dt$ is finite P-almost surely and does not depend on the initial condition $x_1(0)$.*

Proof of Proposition 5.1. Define

$$f_k = \int_{(k-1)\theta}^{k\theta} y(s)\,ds = \int_0^{\theta} y(s + (k-1)\theta)\,ds,$$

$$f_k^n = \sum_{i=0}^{n-1} \frac{\theta}{n}\, y\left(\frac{i\theta}{n} + (k-1)\theta\right).$$

It is clear that $\lim_{n\to\infty}(f_{1+p}^n, \cdots, f_{q+p}^n) = (f_{1+p}, \cdots, f_{q+p})$ P-almost surely for all p, q. From periodicity the repartition of $(f_{1+p}^n, \cdots, f_{q+p}^n)$ does not depend on p. Then the repartition of $(f_{1+p}, \cdots, f_{q+p})$ is independent on p, i.e., the sequence (f_k) is stationary. The result is now a consequence of the ergodic theorem (see [2, Thm. 6.28] or [18, Thm. 3]).

Proof of Proposition 5.2. According to Proposition 3.2 the process

$$x(t) = \int_{-\infty}^t U_{A-BC}(t,s)[f(s) + B(s)v(s)]\,ds + \int_{-\infty}^t U_{A-BC}(t,s)G(s)\,dw(s)$$

is θ-periodic and is a mild solution of (5.1), which satisfies (5.4), and in particular, (5.3) also holds.

If y is another θ-periodic mild solution of (5.1), then for fixed t and for all $s<t$ we have

$$\begin{aligned} E[|x(t)-y(t)|^2] &= E[|U_{A-BC}(t,s)(x(s)-y(s))|^2] \\ &\leqq \alpha^2 \exp\{-2\beta(t-s)\}E[|x(s)-y(s)|^2] \\ &\leqq \gamma_7 \exp(2\beta s) \to 0 \quad \text{as } s \to -\infty, \end{aligned}$$

so that $x(t)=y(t)$ P-almost surely.

Next we can write $x_1(t)=x(t)+z(t)$, where $x(t)$ is the θ-periodic solution of (5.1) and $z(t)=U_{A-BC}(t,0)[x_1(0)-x(0)]$.

If $J_1(u_1)$ is the cost defined in Proposition 5.2 then by a simple computation we have

$$J_1(u_1)=\overline{\lim_{n\to\infty}}\left\{\frac{1}{n}\int_0^{n\theta} y_1(t)\,dt+\frac{1}{n}\int_0^{n\theta} y_2(t)\,dt+\frac{1}{n}\int_0^{n\theta} y_3(t)\,dt\right\}$$

where y_1 is a θ-periodic process with $\sup_t E(|y_1(t)|)<\infty$, $y_2(t)\leqq \gamma_8|z(t)|+\gamma_9|z(t)|^2$, and $y_3(t)\leqq y(t)|z(t)|$, with y θ-periodic and $\sup_t E(|y(t)|^2)<\infty$. From Proposition 5.1 we have that $(1/n)\int_0^{n\theta} y_1(t)\,dt \to^{\text{a.s.}} f$. Also it is easily seen that $(1/n)\int_0^{n\theta} y_2(t)\,dt\to 0$.

Define

$$\begin{aligned} g_k &= \int_{(k-1)\theta}^{k\theta} \exp(-\beta t)y(t)\,dt \leqq \exp\{-\beta(k-1)\theta\}\int_0^{\theta} y(t+(k-1)\theta)\,dt \\ &= \exp\{-\beta(k-1)\theta\}h_k. \end{aligned}$$

By Proposition 5.1 we have that (h_k) is stationary with $E(h_1^2)<\infty$. Since

$$\sum_n P(\exp\{-\beta(n-1)\theta\}h_n>\varepsilon)\leqq \sum_n E(h_1)/\varepsilon\exp(\beta(n-1)\theta)<\infty,$$

it follows that $\exp\{-\beta(n-1)\theta\}h_n\to 0$ P-almost surely, where from

$$\frac{1}{n}\left|\int_0^{n\theta} y_3(t)\,dt\right| \leqq \frac{\gamma_{10}}{n}\sum_{k=1}^{n}\exp\{-\beta(k-1)\theta\}h_k\to 0 \quad P\text{-a.s.}$$

The main result is the following theorem.

THEOREM 5.3. *Assume that the hypotheses of Theorem* 4.2 *hold. Moreover, suppose that*

(a) $D(A(t))=D$ *for all* t *and* $\sup_{0\leqq t\leqq\theta}|A(t)h|<\infty$ *for all* $h\in D$;

(b) 1 *belongs to the resolvent set of* $U_{L^*(Q)}$, *where* Q *is the solution of* (4.1).

Then we have the following assertions:

(i) *The optimal control* $\tilde{u}$ *is given by the feedback law*

$$\tilde{u}=-N^{-1}B^*(Q\tilde{x}+r) \tag{5.6}$$

where r *is the solution of* (4.8).

The cost functional (5.5) *satisfies P-almost surely the equality*

$$\begin{aligned} J(\tilde{u}) = \int_0^{\theta} &[2\langle f(s), r(s)\rangle - \langle N^{-1}(s)B^*(s)r(s), B^*(s)r(s)\rangle \\ &+\operatorname{Tr}(G^*(s)Q(s)G(s)W)]\,ds \end{aligned} \tag{5.7}$$

and the optimal dynamics $\tilde{x}(t)$ corresponding to $\tilde{u}$ is given by

$$\begin{aligned}\tilde{x}(t) &= \int_{-\infty}^{t} U_{L(Q)}(t,s)[f(s) - B(s)N^{-1}(s)B^*(s)r(s)]\,ds \\ &\quad + \int_{-\infty}^{t} U_{L(Q)}(t,s)G(s)\,dw(s).\end{aligned} \tag{5.8}$$

(ii) *Let us assume in addition that* (4.10) *holds. If $u \in U_{ad}$ is such that $\lim_{n\to\infty}(1/n)\int_0^{n\theta}|u(t)-\bar{u}(t)|^2\,dt > 0$ P-almost surely, then P-almost surely there exists n' such that for $n \geqq n'$*

$$\begin{aligned}&\int_0^{n\theta}[\langle M(t)x(t), x(t)\rangle + \langle N(t)u(t), u(t)\rangle]\,dt \\ &\quad > \int_0^{n\theta}[\langle M(t)\tilde{x}(t), \tilde{x}(t)\rangle + \langle N(t)\tilde{u}(t), \tilde{u}(t)\rangle]\,dt\end{aligned}$$

where x is the mild solution of (5.1) *corresponding to u.*

We need the following lemma.

LEMMA 5.4. *Assume the hypotheses of Theorem* 5.3 *are satisfied. Let $u \in U_{ad}$, x be the mild solution of* (5.1) *corresponding to u, and define $\bar{u} = -N^{-1}B^*(Qx+r)$. Then the following identity holds*:

$$\begin{aligned}&\int_0^t [\langle M(s)x(s), x(s)\rangle + \langle N(s)u(s), u(s)\rangle\,ds \\ &\quad = \int_0^t \langle N(s)[u(s)-\bar{u}(s)], u(s)-\bar{u}(s)\rangle\,ds \\ &\qquad + \int_0^t [2\langle f(s), r(s)\rangle - \langle N^{-1}(s)B^*(s)r(s), B^*(s)r(s)\rangle \\ &\qquad + \mathrm{Tr}\,(G^*(s)Q(s)G(s)W)]\,ds \\ &\qquad + \langle Q(0)x(0), x(0)\rangle - \langle Q(t)x(t), x(t)\rangle \\ &\qquad + 2\langle r(t), x(t)\rangle - 2\langle r(0)x(0), x(0)\rangle \\ &\qquad + 2\int_0^t \langle Q(s)x(s) - r(s), G(s)\,dw(s)\rangle.\end{aligned} \tag{5.9}$$

Proof of Lemma 5.4. Let $R_\lambda(t)$ be the resolvent of $A(t)$, and let x_λ be the strong solution of

$$\begin{aligned}dx_\lambda(t) &= \{A(t)x_\lambda(t) + \lambda R_\lambda(t)[B(t)u(t)+f(t)]\}\,dt \\ &\quad + \lambda R_\lambda(t)G(t)\,dw(t), \qquad t \geqq 0, \\ x_\lambda(0) &= \lambda R_\lambda(0)x(0).\end{aligned} \tag{5.10}$$

Let $\bar{u}_\lambda = -N^{-1}B^*(Qx_\lambda + r)$. It is known that for every t

$$x_\lambda(t) \to x(t) \text{ in probability, as } \lambda \to \infty \text{ (see [4], [12], [13]).} \tag{5.11}$$

By using Ito's formula we obtain

$$
\begin{aligned}
&\langle Q(t)x_\lambda(t), x_\lambda(t)\rangle \\
&\quad = \langle Q(0)x_\lambda(0), x_\lambda(0)\rangle \\
&\qquad + \int_0^t \Big\{ \frac{d}{ds}\langle Q(s)x_\lambda(s), x_\lambda(s)\rangle + 2\langle Q(s)x_\lambda(s), A(s)x_\lambda(s)\rangle \\
&\qquad\quad + 2\langle Q(s)x_\lambda(s), \lambda R_\lambda(s)[B(s)u(s)+f(s)]\rangle \\
&\qquad\quad + \mathrm{Tr}\,[G^*(s)\lambda R_\lambda^*(s)Q(s)\lambda R_\lambda(s)Q(s)W]\Big\}\, ds \\
&\qquad + 2\int_0^t \langle Q(s)x_\lambda(s), \lambda R_\lambda(s)G(s)\, dw(s)\rangle
\end{aligned} \tag{5.12}
$$

where from substituting $d/ds\langle Q(s)x_\lambda(s), x_\lambda(s)\rangle$ by (4.12) we deduce

$$
\begin{aligned}
&\int_0^t \langle M(s)x_\lambda(s), x_\lambda(s)\rangle\, ds \\
&\quad = \langle Q(0)x_\lambda(0), x_\lambda(0)\rangle - \langle Q(t)x_\lambda(t), x_\lambda(t)\rangle \\
&\qquad + \int_0^t [\langle QBN^{-1}B^*Qx_\lambda, x_\lambda\rangle + 2\langle Qx_\lambda, \lambda R_\lambda(Bu+f)\rangle \\
&\qquad + \mathrm{Tr}\,(G^*\lambda R_\lambda^* Q\lambda R_\lambda GW)]\, ds + 2\int_0^t \langle Qx_\lambda, \lambda R_\lambda G\, dw\rangle.
\end{aligned} \tag{5.13}
$$

Also we have

$$
\begin{aligned}
\langle Nu, u\rangle &= \langle N(u-\bar{u}_\lambda), u-\bar{u}_\lambda\rangle - 2\langle u, B^*Qx_\lambda\rangle - 2\langle u, B^*r\rangle \\
&\quad - \langle N^{-1}B^*Qx_\lambda, B^*Qx_\lambda\rangle - \langle N^{-1}B^*Qx_\lambda, B^*r\rangle \\
&\quad - \langle N^{-1}B^*r, B^*Qx_\lambda\rangle - \langle N^{-1}B^*r, B^*r\rangle.
\end{aligned} \tag{5.14}
$$

Now from (5.13) and (5.14), we obtain

$$
\begin{aligned}
I_\lambda(t) &= \int_0^t [\langle M(s)x_\lambda(s), x_\lambda(s)\rangle + \langle N(s)u(s), u(s)\rangle]\, ds \\
&= \langle Q(0)x_\lambda(0), x_\lambda(0)\rangle - \langle Q(t)x_\lambda(t), x_\lambda(t)\rangle \\
&\quad + \int_0^t [\langle N(u-\bar{u}_\lambda), u-\bar{u}_\lambda\rangle + 2\langle Qx_\lambda, \lambda R_\lambda(Bu+f)\rangle + \mathrm{Tr}\,(G_\lambda^* R_\lambda^* Q\lambda R_\lambda GW) \\
&\quad - 2\langle u, B^*Qx_\lambda\rangle - 2\langle u, B^*r\rangle - 2\langle N^{-1}B^*Qx_\lambda, B^*r\rangle - \langle N^{-1}B^*r, B^*r\rangle]\, ds \\
&\quad + 2\int_0^t \langle Qx_\lambda, \lambda R_\lambda G dw\rangle.
\end{aligned} \tag{5.15}
$$

Next, by using Ito's formula for $\langle r(t)x_\lambda(t), x_\lambda(t)\rangle$ and then substituting r' by (4.8), we arrive at

$$
\begin{aligned}
&\int_0^t \langle QBN^{-1}B^*r, x_\lambda\rangle\, ds \\
&\quad = \langle r(t), x_\lambda(t)\rangle - \langle r(0), x_\lambda(0)\rangle \\
&\qquad + \langle Q(t)f(t), x_\lambda(t)\rangle - \langle r(t), \lambda R_\lambda(t)[B(t)u(t)+f(t)]\rangle \\
&\qquad + \int_0^t \langle r, \lambda R_\lambda G\, dw\rangle.
\end{aligned} \tag{5.16}
$$

Substituting (5.16) in (5.15), we get

$$
\begin{aligned}
I_\lambda(t) = & \int_0^t \langle N(u-\bar{u}_\lambda), u-\bar{u}_\lambda\rangle\, ds \\
& + \int_0^t [\mathrm{Tr}\,(G^*\lambda R_\lambda^* Q \lambda R_\lambda G W) - \langle N^{-1}B^*r, B^*r\rangle + 2\langle r, \lambda R_\lambda f\rangle]\, ds \\
& + \langle Q(0)x_\lambda(0), x_\lambda(0)\rangle - \langle Q(t)x_\lambda(t), x_\lambda(t)\rangle \\
& + 2\langle r(t), x_\lambda(t)\rangle - 2\langle r(0), x_\lambda(0)\rangle \\
& + \int_0^t [2\langle Qx_\lambda, \lambda R_\lambda(Bu+f)\rangle - 2\langle u, B^*Qx_\lambda\rangle - 2(u, B^*r\rangle \\
& - 2\langle Qf, x_\lambda\rangle + 2\langle r, \lambda R_\lambda Bu\rangle]\, ds \\
& + 2\int_0^t \langle Qx_\lambda - r, \lambda R_\lambda G\, dw\rangle.
\end{aligned} \tag{5.17}
$$

Finally, letting $\lambda \to \infty$ and utilizing (5.11), we obtain (5.9).

Proof of Theorem 5.3. (i) If we take $u = \tilde{u}$ $(= \tilde{u})$, $t = n\theta$, we divide by n in (5.9), and then we apply Lemma 3.3; we obtain (5.7). The equality (5.8) follows from Proposition 3.2.

(ii) Part (ii) is a consequence of (5.9) and of Lemma 3.3.

Remark 5.1. The result in (5.7) is of the "limit theorem" type as mentioned in [15] and $\tilde{u}$ can be regarded as optimal with respect to the law of the large numbers.

Example. Consider the stochastic parabolic equation formally described by

$$
\begin{aligned}
& dx(t,z) = \left\{\left[\frac{\partial^2}{\partial z^2} - \varphi(t)\right] x(t,z) + u(t,z) + f(t,z)\right\} dt + G(t,z)\, dw(t), \\
& x(t,0) = x(t,1) = 0
\end{aligned} \tag{5.18}
$$

where φ, f, G are continuous in (t,z), θ-periodic in t, $\varphi \geqq 0$ and w is a real Wiener process.

Let $K = R$, $H = U = L^2((0,1), R)$, $A(t) = \partial^2/\partial z^2 - \varphi(t)$, $D(A(t)) = D = \{h \in H;\ \dot{h}, \ddot{h} \in H,\ h(0) = h(1) = 0\}$.

$A(t)$ generates a strong evolution operator U with $\|U(t,s)\| \leqq 1$ for all $S \leqq t$. The evolution equation corresponding to (5.18) is

$$
dx(t) = [A(t)x(t) + u(t) + f(t)]\, dt + G(t)\, dw(t). \tag{5.19}
$$

THEOREM 5.5. *The cost functional*

$$
J(u) = \overline{\lim_{n\to\infty}} \frac{1}{n} \int_0^{n\theta} [|x(t)|_H^2 + |u(t)|_H^2]\, dt \tag{5.20}
$$

has P-almost surely the minimum value

$$
J(\tilde{u}) = \int_0^\theta [2\langle r(s), f(s)\rangle - |r(s)|_H^2 + \mathrm{Tr}\,(G^*(s)Q(s)G(s))]\, ds \tag{5.21}
$$

corresponding to the optimal control

$$
\tilde{u}(t) = -Q(t)\tilde{x}(t) - r(t) \tag{5.22}
$$

where Q, r are the θ-periodic solutions of

$$\frac{d}{dt}\langle Qh, h\rangle + 2\langle Ah, Qh\rangle - |Qh|_H^2 + |h|_H^2 = 0, \qquad h \in D, \tag{5.23}$$

$$r' + (A - Q)r + Qf = 0, \tag{5.24}$$

and $\tilde{x}$ is the θ-periodic process

$$\tilde{x}(t) = \int_{-\infty}^{t} U_{A-Q}(t, s)[f(s) - r(s)]\, ds + \int_{-\infty}^{t} U_{A-Q}(t, s)G(s)\, dw(s). \tag{5.25}$$

Acknowledgments. I thank the referees for many valuable comments and suggestions and Professor A. Halanay for helpful discussions.

REFERENCES

[1] L. ARNOLD, R. F. CURTAIN, AND P. KOTELENEZ, *Nonlinear stochastic evolution equations in Hilbert spaces*, Report no. 17, Forschungsschwerpunkt Dynamische Systeme, Universitat Bremen, Bremen, FRG, 1980.

[2] L. BREIMAN, *Probability*, Addison-Wesley, Reading, MA, 1968.

[3] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite-dimensional Riccati equation for systems defined by evolution operators*, SIAM J. Control, 14 (1976), pp. 951-983.

[4] R. F. CURTAIN, *Markov processes generated by linear stochastic evolution equations*, Stochastics, 5 (1981), pp. 135-165.

[5] G. DA PRATO, *Synthesis of optimal control for an infinite dimensional periodic problem*, SIAM J. Control Optim., 25 (1987), pp. 706-714.

[6] G. DA PRATO AND A. ICHIKAWA, *Optimal control of linear systems with almost periodic inputs*, SIAM J. Control Optim., 25 (1987), pp. 1007-1019.

[7] ———, *Quadratic control for linear periodic systems*, Appl. Math. Optim., 18 (1988), pp. 39-66.

[8] ———, *Filtering and control of linear periodic systems*, submitted for publication.

[9] T. DATKO, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428-445.

[10] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537-565.

[11] A. HALANAY, C. TUDOR, AND T. MOROZAN, *Tracking almost periodic signals under white noise perturbations*, Stochastics, 21 (1987), pp. 287-301.

[12] A. ICHIKAWA, *Dynamic programming approach to stochastic evolution equations*, SIAM J. Control Optim., 17 (1979), pp. 152-174.

[13] ———, *Stability of semilinear stochastic evolution equations*, J. Math. Anal., 90 (1982), pp. 12-44.

[14] ———, *Equivalence of L_p stability and exponential stability for a class of nonlinear semigroups*, Nonlinear Anal., 8 (1984), pp. 805-815.

[15] P. MANDL, *Limit theorems of probability theory and optimality in linear controlled systems with quadratic cost*, in Stochastic Systems, Lecture Notes in Control and Information Sciences 96, H. J. Engelbert and W. Schmidt, eds., Springer-Verlag, Berlin, New York, 1987.

[16] M. METIVIER AND J. PELLAUMAIL, *Stochastic Integration*, Academic Press, New York, 1980.

[17] T. MOROZAN, *Bounded and periodic solutions of affine stochastic differential equations*, Studii si Cercetari Matematice, 38 (1986), pp. 523-527.

[18] A. N. SHIRYAYEV, *Probability*, Springer-Verlag, Berlin, New York, 1984.

[19] H. TANABE, *Equations of Evolution*, Pitman, London, San Francisco, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 265-293, March 1990

QUADRATIC OPTIMIZATION FOR INFINITE-DIMENSIONAL LINEAR DIFFERENTIAL DIFFERENCE TYPE SYSTEMS: SYNTHESES VIA THE FREDHOLM EQUATION*

E. BRUCE LEE† AND YUNCHENG YOU‡

Abstract. A closed-loop solution for quadratic optimization is presented for control systems whose model is given by infinite-dimensional linear differential difference equations:

$$\frac{dx(t)}{dt} = Ax(t) + \sum_{i=1}^{n} A_i x(t-h_i) + Bu(t) + \sum_{j=1}^{m} B_j u(t-r_j).$$

A nonsemigroup evolution formula for the fundamental solution is given first. The system is then reduced to a Volterra integral system. By the semicausality approach, a new optimality principle is established, and the closed-loop optimal control is given by real-time state feedback plus retarded state integral feedback. The feedback operator is determined by solving a linear Fredholm integral equation. The two crucial methods, semicausal dynamical optimization for Volterra integral systems and nonsemigroup evolution property, are brought together to generalize the syntheses results to infinite-dimensional cases.

Key words. optimal control, infinite-dimensional linear system, differential difference equation, semicausality, Fredholm equation, synthesis solution

AMS(MOS) subject classifications. 49B22, 34K35, 45B05, 93C25

1. Optimal control formulation. The quadratic optimization theory of finite-dimensional linear systems with delay in state and/or control variables has been developed by many authors via several different approaches (cf. [1]-[4], [6]-[9], [11]-[15], [18]-[23] and the references therein). Solving a Riccati type equation is a standard way to provide the feedback operator when delays are not involved in the quadratic formulation [10], and also in the finite-dimensional delay cases [8]. The method used to provide the feedback operator for the infinite-dimensional delay type system that is given here is based on the Fredholm resolvent theory. Schumitzky [22] has described the connection between the Riccati equations and the Fredholm resolvent theory. Manitius [20] has used the Fredholm resolvent theory to provide the feedback operator for quadratic optimization for finite-dimensional systems with delays. You [24]-[26] has now generalized this approach to certain infinite-dimensional systems (partial differential equation (PDE) models). Background results on infinite-dimensional systems can be found in [5]. Delfour [8] gives general results in the finite-dimensional case along with associated coupled differential Riccati equations.

However, the methods based on some kind of Riccati equation are quite restrictive for optimization with delayed output appearing in the quadratic criteria (cf. [18]), and also in the infinite-dimensional case there are complications in deriving appropriate Riccati equations, as well as a lack of effective analytic or numerical means to solve the infinite-dimensional differential or integral Riccati equations that will be inherently of the same dimension as the state space. In view of this fact, we have introduced [18] a new approach based on the concepts of a truncation operator and semicausal

* Received by the editors March 16, 1987; accepted for publication (in revised form) March 24, 1989. This research was supported by National Science Foundation grants DMS-8607687 and DMS-8722402.

† Department of Electrical Engineering, Center for Control Science and Dynamical Systems, University of Minnesota, Minneapolis, Minnesota 55455.

‡ School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

trajectory to obtain a synthesis with time delays both in state and in output variables; the latter appears in the cost functional. We have also applied this approach [16], [17] to optimization connected with Volterra integral type systems and two-dimensional systems.

In this paper, this semicausality approach, combined with the nonsemigroup evolution formula of the fundamental solution, is further developed for the infinite-dimensional differential difference type linear systems with delays in state and control variables. A closed-loop synthesis will be presented with the feedback operator determined by solving a linear Fredholm integral equation, which possesses the same dimension as the control variable space (usually finite even though the state space is infinite-dimensional). Therefore, the obtained results will be more applicable theoretically and computationally.

Let X and U be real Hilbert spaces. $T>0$ is finite and fixed. Consider a linear system as represented by infinite-dimensional differential-difference equations:

$$\frac{dx(t)}{dt}=Ax(t)+\sum_{i=1}^{n}A_ix(t-h_i)+Bu(t)+\sum_{j=1}^{m}B_ju(t-r_j),\qquad t\geqq 0, \tag{1.1}$$

$$x(\theta)=\phi(\theta),\quad -h\leqq\theta<0,\quad x(0)=x_0,\quad u(\xi)=\psi(\xi),\quad -r\leqq\xi\leqq 0, \tag{1.2}$$

where $0<h_1<h_2<\cdots<h_n$ and $0<r_1<r_2<\cdots<r_m$ with $h=h_n$ and $r=r_m$, $0<\max(h, r)\ll T$. Denote by $Z=X\times L^2(-h, 0; X)\times L^2(-r, 0; U)$ and assume that initial data $(x_0, \phi, \psi)\in Z$ is arbitrarily given.

Assume that $A:\mathcal{D}(A)(\subset X)\to X$ is a densely defined and closed operator that generates a C_0 semigroup $e^{At}(t\geqq 0)$, besides $A_i\in\mathcal{L}(X)$, $i=1,\cdots,n$, $B\in\mathcal{L}(U; X)$, and $B_j\in\mathcal{L}(U; X)$, $j=1,\cdots,m$.

The state function is the mild solution of (1.1) with the initial data (1.2), i.e.,

$$x(t)=\begin{cases}\phi(t), & -h\leqq t<0,\\ e^{At}x_0+\displaystyle\int_0^t e^{A(t-s)}\left\{\sum_{i=1}^{n}A_ix(s-h_i)+Bu(s)+\sum_{j=1}^{m}B_ju(s-r_j)\right\}ds, & t\geqq 0.\end{cases} \tag{1.3}$$

The admissible control set is given by

$$u(\cdot)\in\mathcal{U}=L^2(0, T; U). \tag{1.4}$$

Set a quadratic cost functional to be

$$J(u)=\langle Q_Tx(T), x(T)\rangle+\int_0^T[\langle Qx(t), x(t)\rangle+\langle Ru(t), u(t)\rangle]\,dt \tag{1.5}$$

where Q_T and $Q\in\mathcal{L}(X)$ are nonnegative and $R\in\mathcal{L}(U)$ is coercively positive.

Optimization involves finding a closed-loop optimal control that minimizes $J(u)$ over $\mathcal{U}$.

Before considering the question of optimal control, it is necessary to make some preparation for the appropriate reformulation of the system in this infinite-dimensional case, as suggested by the following results.

LEMMA 1. *The homogeneous differential-difference equation*

$$\frac{dx(t)}{dt}=Ax(t)+\sum_{i=1}^{n}A_ix(t-h_i),\qquad t\geqq 0, \tag{1.6}$$

$$x(0)=x_0,\quad x(\theta)=\phi(\theta),\quad -h\leqq\theta<0$$

admits a unique mild solution $x(t)$, $t \geqq 0$, *for any given* $(x_0, \phi) \in M^2 = X \times L^2(-h, 0; X)$. *Define* $V(t): (x_0, \phi) \to (x(t), x_t)$, *where* $x_t(\theta) = x(t+\theta)$, *for* $-h \leqq \theta \leqq 0$; *then* $V(t)$ $(t \geqq 0)$ *is a* C_0 *semigroup of bounded linear operators on the* M^2 *space, which is called the solution semigroup associated with* (1.6).

The proof of this lemma is given in Theorem 1 of [25].

Define

$$W^2 = \left\{ \begin{pmatrix} \phi_0 \\ \phi \end{pmatrix} \in M^2 : \phi_0 = \phi(0) \in X, \quad \phi \in AC([-h, 0]; X) \text{ and } \frac{d\phi}{d\theta} \in L^2(-h, 0; X) \right\}$$

with the M^2-induced topology, where AC means strong absolute continuity and $d\phi/d\theta$ is the strong derivative of ϕ.

LEMMA 2. *Let* $\mathcal{A}$ *be the infinitesimal generator of the solution semigroup* $V(t)$ *associated with* (1.6). *Then,* $\mathcal{D}(\mathcal{A}) = \hat{W}^2$ *and*

$$\mathcal{A}\begin{pmatrix} \phi_0 \\ \phi \end{pmatrix} = \begin{pmatrix} A\phi_0 + \sum_{i=1}^{n} A_i \phi(-h_i) \\ E_\phi \end{pmatrix} \quad \forall \begin{pmatrix} \phi_0 \\ \phi \end{pmatrix} \in \mathcal{D}(\mathcal{A}), \tag{1.7}$$

where

$$\hat{W}^2 = \left\{ \begin{pmatrix} \phi_0 \\ \phi \end{pmatrix} \in W^2 : \phi_0 = \phi(0) \in \mathcal{D}(A) \right\},$$

and

$$E_\phi = \frac{d\phi}{d\theta} \quad \text{with } \mathcal{D}(E) = P_{L^2} \hat{W}^2,$$

where $P_{L^2}: M^2 \to L^2(-h, 0; X)$ *is an orthogonal projection.*

The proof of this lemma is also given in Theorem 2 of [25].

By the definition, we see that the mild solution of (1.6) is expressed by

$$x(t) = P_X V(t) \begin{pmatrix} x_0 \\ \phi \end{pmatrix}, \qquad t \geqq 0, \tag{1.8}$$

where $P_X : M^2 \to X$ is an orthogonal projection.

Define the fundamental solution $G(t): [0, \infty) \to \mathcal{L}(X)$ to be the mild solution of the following operator equation:

$$\frac{dG(t)}{dt} = AG(t) + \sum_{i=1}^{n} A_i G(t - h_i), \qquad t \geqq 0, \tag{1.9}$$

$$G(0) = I \quad \text{and} \quad G(t) = 0, \quad t < 0.$$

This mild solution $G(t)$ $(t \geqq 0)$ exists uniquely and is given by

$$G(t) = \begin{cases} e^{At} + \displaystyle\int_0^t e^{A(t-s)} \sum_{i=1}^{n} A_i G(s - h_i)\, ds, & t \geqq 0, \\ 0, & t < 0. \end{cases} \tag{1.10}$$

Since there are multidelays involved in (1.9), the fundamental solution $G(t)$ does not possess the semigroup property, as in the single delay case [8]. However, we need a precise evolutionary property of $G(t)$, which will be used later in the closed-loop syntheses.

LEMMA 3. *Let $G(t)$ be the fundamental solution. Assume that $g(\cdot) \in C^1([0, b]; X)$ with null extension outside $[0, b]$. Then, for any $t \geqq 0$,*

$$\int_0^t G(t-\sigma)g(\sigma)\, d\sigma \in \mathcal{D}(A), \tag{1.11}$$

where $\mathcal{D}(A)$ is the domain of the generator A.

Proof. It is known [5], [23] that for any strongly continuous differentiable $g(\cdot):[0, b] \to X$ (with null extension outside), $\int_0^t e^{A(t-s)}g(s)\, ds \in \mathcal{D}(A)$, and

$$A\int_0^t e^{A(t-s)}g(s)\, ds = e^{At}g(0) - g(t) + \int_0^t e^{A(t-s)}\dot{g}(s)\, ds, \qquad t \geqq 0. \tag{1.12}$$

Since $G(t)$ is given by (1.10), it is enough to show that

$$\int_0^t \int_0^{t-\sigma} e^{A(t-\sigma-s)}A_i G(s-h_i)g(\sigma)\, ds\, d\sigma \in \mathcal{D}(A), \qquad i = 1, \cdots, n. \tag{1.13}$$

Indeed,

$$\begin{aligned} &\int_0^t \int_0^{t-\sigma} e^{A(t-\sigma-s)}A_i G(s-h_i)g(\sigma)\, ds\, d\sigma \\ &= \int_0^t \int_\sigma^t e^{A(t-\xi)}A_i G(\xi-\sigma-h_i)g(\sigma)\, d\xi\, d\sigma \\ &= \int_0^t e^{A(t-\xi)} \int_0^\xi A_i G(\xi-\sigma-h_i)g(\sigma)\, d\sigma\, d\xi. \end{aligned} \tag{1.14}$$

Now we see that $\beta(\xi) \triangleq \int_0^\xi A_i G(\xi-\sigma-h_i)g(\sigma)\, d\sigma = \int_0^\xi A_i G(\eta-h_i)g(\xi-\eta)\, d\eta$ is strongly continuously differentiable (with possible discontinuity in connection with null extension), so that (1.14) implies that (1.13) is true. Moreover, by (1.12) we obtain

$$\begin{aligned} &A\int_0^t \int_0^{t-\sigma} e^{A(t-\sigma-s)}A_i G(s-h_i)g(\sigma)\, ds\, d\sigma \\ &= \int_0^t e^{A(t-\xi)}\left(\frac{d}{d\xi}\int_0^\xi A_i G(\xi-\sigma-h_i)g(\sigma)\, d\sigma\right) d\xi \\ &\quad - A_i \int_0^t G(t-\sigma-h_i)g(\sigma)\, d\sigma. \end{aligned} \tag{1.15}$$

LEMMA 4. *Under the same assumptions as in Lemma 3.*

$$\begin{aligned} \frac{d}{dt}\int_0^t G(t-\sigma)g(\sigma)\, d\sigma &= A\int_0^t G(t-\sigma)g(\sigma)\, d\sigma \\ &\quad + \sum_{i=1}^n A_i \int_0^t G(t-\sigma-h_i)g(\sigma)\, d\sigma + g(t) \end{aligned} \tag{1.16}$$

for $t \geqq 0$ except at the connection point $t = b$ with null extension, where $g(t-0)$ appears for the left derivative.

LEMMA 5. *For any $(x_0, \phi) \in M^2$, the unique mild solution of (1.6) is given by*

$$x(t) = G(t)x_0 + \int_0^t G(t-\sigma) \sum_{i=1}^n A_i \phi(\sigma - h_i)\, d\sigma, \qquad t \geqq 0 \tag{1.17}$$

where it is a convention that $\phi(t) \equiv 0$ whenever $t > 0$.

Proof. First assume that $(x_0, \phi) \in \mathcal{D}(A) \times C^1([-h, 0]; X)$. Let

$$g(\sigma) = \sum_{i=1}^{n} A_i \phi(\sigma - h_i), \qquad \sigma \geqq 0. \tag{1.18}$$

Then we see that $g(\cdot)$ is piecewise strongly continuously differentiable as follows:

$$g(\sigma) = \begin{cases} \sum_{i=1}^{n} A_i \phi(\sigma - h_i), & \sigma \in [0, h_1], \\ \sum_{i=2}^{n} A_i \phi(\sigma - h_i), & \sigma \in (h_1, h_2], \\ \cdots & \cdots \\ A_n \phi(\sigma - h_n), & \sigma \in (h_{n-1}, h_n], \\ 0, & \sigma > h_n. \end{cases} \tag{1.19}$$

Now let

$$y(t) = \begin{cases} G(t)x_0 + \int_0^t G(t-\sigma) \sum_{i=1}^{n} A_i \phi(\sigma - h_i)\, d\sigma, & t \geqq 0, \\ \phi(t), & t < 0. \end{cases} \tag{1.20}$$

By Lemma 4, $y(t)$ is piecewise strongly continuous differentiable on $[0, \infty)$ except for the finite points $t = h_1, \cdots, h_n$, where the left and right strong derivatives still exist. Furthermore, with $g(\cdot)$ defined by (1.18) or (1.19), it follows that for $t \geqq 0$,

$$\begin{aligned} \frac{d}{dt} y(t) &= \frac{d}{dt} G(t)x_0 + \frac{d}{dt} \int_0^t G(t-\sigma) g(\sigma)\, d\sigma \\ &\doteq AG(t)x_0 + \sum_{i=1}^{n} A_i G(t-h_i)x_0 + A \int_0^t G(t-\sigma) g(\sigma)\, d\sigma \\ &\qquad + \sum_{i=1}^{n} A_i \int_0^t G(t-\sigma-h_i) g(\sigma)\, d\sigma + g(t) \\ &= Ay(t) + \left\{ \sum_{i=1}^{n} A_i G(t-h_i)x_0 + \sum_{i=1}^{n} A_i \int_0^t G(t-\sigma-h_i) g(\sigma)\, d\sigma + g(t) \right\} \\ &= Ay(t) + \begin{cases} \sum_{i=1}^{n} A_i \phi(t-h_i), & \text{for } t \in [0, h_1], \\ A_1(G(t-h_1)x_0 + \int_0^t G(t-\sigma-h_1) g(\sigma)\, d\sigma) + \sum_{i=2}^{n} A_i \phi(t-h_1), & \\ & \text{for } t \in (h_1, h_2], \\ \cdots \qquad\qquad \cdots & \\ \sum_{i=1}^{n-1} A_i(G(t-h_i)x_0 + \int_0^t G(t-\sigma-h_i) g(\sigma)\, d\sigma) + A_n \phi(t-h_n), & \\ & \text{for } t \in (h_{n-1}, h_n], \\ \sum_{i=1}^{n} A_i(G(t-h_i)x_0 + \int_0^t G(t-\sigma-h_i) g(\sigma)\, d\sigma), & \text{for } t \in (h_n, \infty), \end{cases} \\ &= Ay(t) + \sum_{i=1}^{n} A_i y(t-h_i) \end{aligned} \tag{1.21}$$

where the first type of discontinuity exists at the points $h_1, \cdots, h_n$.

On the other hand, it can be shown that the mild solution $x(t)$ of (1.6) with the same initial data $(x_0, \phi) \in \mathcal{D}(A) \times C^1([-h, 0]; X)$ is also piecewise strongly continuously differentiable and

$$\frac{dx(t)}{dt} = Ax(t) + \sum_{i=1}^{n} A_i x(t-h_i) \tag{1.22}$$

for $t \geqq 0$ except at $h_1, \cdots, h_n$. Thus $z(t) = x(t) - y(t)$, $t \in [-h, \infty)$, is a strongly absolutely continuous function with its strong derivative $dz/dt = 0$ almost everywhere. This implies that $z(t) \equiv 0$, so that (1.17) is valid for (x_0, ϕ) in $\mathcal{D}(A) \times C^1([-h, 0]; X)$.

Finally, since $\mathcal{D}(A) \times C^1([-h, 0]; X)$ is dense in M^2, we can always use some sequence $\{(x_0^{(k)}, \phi^{(k)})\}_{k=1}^{\infty} \subset \mathcal{D}(A) \times C^1([-h, 0]; X)$ to approximate a given $(x_0, \phi) \in M^2$ so that corresponding sequences $\{x^{(k)}(\cdot)\}_{k=1}^{\infty}$ and $\{y^{(k)}(\cdot)\}_{k=1}^{\infty}$ will converge to $x(\cdot)$ and $y(\cdot)$ with the same (x_0, ϕ) as the initial data, respectively. That is, $\lim_{k\to\infty} y^{(k)}(t) = y(t)$ and $\lim_{k\to\infty} x^{(k)}(t) = x(t)$ for $t \geqq 0$. Since $x^{(k)}(t) \equiv y^{(k)}(t)$ for all $t \geqq 0$, it follows that (1.17) is valid for all $(x_0, \phi) \in M^2$. □

THEOREM 1. *Let $G(t)$ be the fundamental solution. Then it satisfies the following relation*:

$$G(t+s) = G(t)G(s) + \sum_{i=1}^{n} \int_0^{\min(t,h_i)} G(t-\sigma) A_i G(s+\sigma-h_i)\, d\sigma \quad \text{for } t \geqq 0,\ s \geqq 0. \tag{1.23}$$

Proof. By Lemma 5 and the definition of the solution semigroup $V(t)$ associated with (1.6), it follows that

$$\begin{aligned} V(t)\begin{pmatrix} x_0 \\ \phi \end{pmatrix} &= \begin{pmatrix} G(t)x_0 + \displaystyle\int_0^t G(t-\sigma) \sum_{i=1}^{n} A_i \phi(\sigma - h_i)\, d\sigma \\ G(t+\theta)x_0 + \displaystyle\int_0^{t+\theta} G(t+\theta-\sigma) \sum_{i=1}^{n} A_i \phi(\sigma-h_i)\, d\sigma, \quad \theta \in [-h, 0] \end{pmatrix} \\ &= \begin{pmatrix} G(t)x_0 + \displaystyle\sum_{i=1}^{n} \int_0^{\min(t,h_i)} G(t-\sigma) A_i \phi(\sigma-h_i)\, d\sigma \\ G(t+\theta)x_0 + \displaystyle\sum_{i=1}^{n} \int_0^{\min(t+\theta,h_i)} G(t+\theta-\sigma) A_i \phi(\sigma-h_i)\, d\sigma, \quad \theta \in [-h, 0] \end{pmatrix}, \qquad t \geqq 0, \end{aligned} \tag{1.24}$$

where $G(t)$ is the described fundamental solution. In particular,

$$V(t)\begin{pmatrix} x_0 \\ 0 \end{pmatrix} = \begin{pmatrix} G(t)x_0 \\ G(t+\theta)x_0, \quad \theta \in [-h, 0] \end{pmatrix}, \qquad t \geqq 0. \tag{1.25}$$

Now for $t \geqq 0$ and $s \geqq 0$, it follows that

$$\begin{aligned} V(t+s)\begin{pmatrix} x_0 \\ 0 \end{pmatrix} &= \begin{pmatrix} G(t+s)x_0 \\ G(t+s+\theta)x_0, \quad \theta \in [-h, 0] \end{pmatrix} \\ &= V(t)V(s)\begin{pmatrix} x_0 \\ 0 \end{pmatrix} = V(t)\begin{pmatrix} G(s)x_0 \\ G(s+\theta)x_0, \quad \theta \in [-h, 0] \end{pmatrix} \end{aligned} \tag{1.26}$$

$$= \begin{pmatrix} G(t)G(s)x_0 + \sum_{i=1}^{n} \int_0^{\min(t,h_i)} G(t-\sigma)A_iG(s+\sigma-h_i)\,d\sigma\, x_0 \\ G(t+\theta)G(s)x_0 + \sum_{i=1}^{n} \int_0^{\min(t+\theta,h_i)} G(t+\theta-\sigma)A_iG(s+\sigma-h_i)\,d\sigma\, x_0, \quad \theta\in[-h,0] \end{pmatrix}$$

for any $x_0 \in X$.

The equality (1.26) indicates that (1.23) holds. □

For the original linear control system (1.1), (1.2), by an argument similar to that in Lemma 5, it can be proved that the state function of (1.1), (1.2) is expressed by

$$\begin{aligned} x(t) &= G(t)x_0 + \int_0^t G(t-\sigma) \sum_{i=1}^{n} A_i\phi(\sigma-h_i)\,d\sigma \\ &\quad + \int_0^t G(t-\sigma)Bu(\sigma)\,d\sigma + \int_0^t G(t-\sigma) \sum_{j=1}^{m} B_j u(\sigma-r_j)\,d\sigma, \qquad t\in[0,T]. \end{aligned} \tag{1.27}$$

Here and later it will always be a convention that

$$\begin{aligned} &\phi(t)=0 \quad \text{whenever } t>0, \\ &\psi(t)=0 \quad \text{whenever } t>0, \\ &G(t)=0 \quad \text{whenever } t<0. \end{aligned} \tag{1.28}$$

By this convention, the Volterra integrals with time-delayed control can be written as

$$\begin{aligned} \int_0^t G(t-\sigma)B_ju(\sigma-r_j)\,d\sigma &= \int_0^{r_j} G(t-\sigma)B_j\psi(\sigma-r_j)\,d\sigma \\ &\quad + \int_{r_j}^t G(t-\sigma)B_ju(\sigma-r_j)\,d\sigma \\ &= \int_0^t G(t-\sigma)B_j\psi(\sigma-r_j)\,d\sigma + \int_0^t G(t-r_j-\sigma)B_ju(\sigma)\,d\sigma \\ &\quad \text{for } t\geqq r_j, \quad j=1,\cdots,m. \end{aligned} \tag{1.29}$$

If $t<r_j$, then the last term $\int_0^t G(t-r_j-\sigma)B_ju(\sigma)\,d\sigma$ vanishes by (1.28).

Therefore, the state function $x(t)$ is expressed by

$$x(t) = f_{(x_0,\phi,\psi)}(t) + \int_0^t G(t-\sigma)Bu(\sigma)\,d\sigma + \sum_{j=1}^{m} \int_0^t G(t-r_j-\sigma)B_ju(\sigma)\,d\sigma, \tag{1.30}$$

where

$$f_{(x_0,\phi,\psi)}(t) = G(t)x_0 + \int_0^t G(t-\sigma)\left\{ \sum_{i=1}^{n} A_i\phi(\sigma-h_i) + \sum_{j=1}^{m} B_j\psi(\sigma-r_j) \right\} d\sigma. \tag{1.31}$$

Define

$$F(t) = G(t)B + G(t-r_1)B_1 + \cdots + G(t-r_m)B_m, \qquad t\geqq 0; \tag{1.32}$$

then the state function is expressed by the following formula:

$$x(t) = f_{(x_0,\phi,\psi)}(t) + \int_0^t F(t-\sigma)u(\sigma)\,d\sigma, \qquad t\in[0,T], \tag{1.33}$$

where $f_{(x_0,\phi,\psi)}(t)$ and $F(t)$ are defined by (1.31) and (1.32), respectively.

Therefore, the optimal control formulation is to find a closed-loop control function u that minimizes $J(u)$ over $u \in \mathcal{U}$, where $J(u)$ is defined by (1.5) and the state function is given by (1.33).

2. Open-loop optimal control. Denote by $\mathcal{X} = L^2(0, T; X)$ and $C = C([0, T]; X)$. Define operators $\Gamma \in \mathcal{L}(\mathcal{U}; \mathcal{X})$ and $\Gamma_T \in \mathcal{L}(\mathcal{U}; X)$ by

$$(\Gamma u)(t) = \int_0^t F(t-\sigma)u(\sigma)\, d\sigma, \qquad t \in [0, T], \tag{2.1}$$

$$\Gamma_T u = \int_0^T F(T-\sigma)u(\sigma)\, d\sigma. \tag{2.2}$$

Then the state (1.33) can be written as

$$x(\cdot) = f_{(x_0,\phi,\psi)} + \Gamma u, \tag{2.3}$$

with the final value

$$x(T) = f_{(x_0,\phi,\psi)}(T) + \Gamma_T u. \tag{2.4}$$

Substitution of (2.3) and (2.4) into the cost functional (1.5) yields

$$\begin{aligned} J(u) &= \langle Ru, u\rangle_{\mathcal{U}} + \langle Q(\Gamma u + f_{(x_0,\phi,\psi)}), \Gamma u + f_{(x_0,\phi,\psi)}\rangle_{\mathcal{X}} \\ &\quad + \langle Q_T(\Gamma_T u + f_{(x_0,\phi,\psi)}(T)), \Gamma_T u + f_{(x_0,\phi,\psi)}(T)\rangle_X \\ &= \langle \Phi u, u\rangle_{\mathcal{U}} + 2\langle \Gamma^* Q f_{(x_0,\phi,\psi)} + \Gamma_T^* Q_T f_{(x_0,\phi,\psi)}(T), u\rangle_{\mathcal{U}} + \text{const.}\,(x_0, \phi, \psi), \end{aligned} \tag{2.5}$$

where the operator $\Phi \in \mathcal{L}(\mathcal{U})$ is defined by

$$\Phi = RI_{\mathcal{U}} + \Gamma^* Q \Gamma + \Gamma_T^* Q_T \Gamma_T, \tag{2.6}$$

and Φ is coercively positive, and here the same notation Q is used to denote the operator $\tilde{Q} \in \mathcal{L}(\mathcal{X})$ in the sense that

$$(\tilde{Q}y)(t) = Qy(t) \quad \forall y \in \mathcal{X},$$

and const. (x_0, ϕ, ψ) is a constant determined by x_0, ϕ, ψ, i.e.,

$$\text{const.}\,(x_0, \phi, \psi) = \langle Qf_{(x_0,\phi,\psi)}, f_{(x_0,\phi,\psi)}\rangle_{\mathcal{X}} + \langle Q_T f_{(x_0,\phi,\psi)}(T), f_{(x_0,\phi,\psi)}(T)\rangle.$$

THEOREM 2. *For any given $(x_0, \phi, \psi) \in Z \triangleq X \times L^2(-h, 0; X) \times L^2(-r, 0; U)$, there exists a unique optimal control. The control process $(u(\cdot), x(\cdot))$ is optimal if and only if it satisfies the following relation*:

$$u(t) = -R^{-1}\left[F^*(T-t)Q_T x(T) + \int_t^T F^*(\sigma - t)Qx(\sigma)\, d\sigma\right], \qquad t \in [0, T]. \tag{2.7}$$

Proof. From (2.5) with the dominant operator Φ being coercively positive, we obtain that there exists a unique optimal control $u(\cdot) \in \mathcal{U}$ that minimizes $J(u)$ over $\mathcal{U}$. Moreover, $u(\cdot)$ is optimal if and only if

$$\Phi u = -(\Gamma^* Q f_{(x_0,\phi,\psi)} + \Gamma_T^* Q_T f_{(x_0,\phi,\psi)}(T)), \tag{2.8}$$

or equivalently

$$Ru = -\Gamma^* Q x(\cdot) - \Gamma_T^* Q_T x(T). \tag{2.9}$$

Here the adjoint operators Γ^* and Γ_T^* can be given explicitly by

$$(\Gamma^* y)(t) = \int_t^T F^*(\sigma - t) y(\sigma)\, d\sigma, \qquad t \in [0, T] \quad \forall y \in \mathscr{X},$$
$$(\Gamma_T^* y_1)(t) = F^*(T-t) y_1, \qquad t \in [0, T] \quad \forall y_1 \in X. \tag{2.10}$$

Substitute (2.10) into (2.9); then it follows that (2.7) is valid. □

COROLLARY 1. *For any given $(x_0, \phi, \psi) \in Z$, the optimal control $u(\cdot)$ is the unique solution in $\mathscr{U}$ of the following open-loop equation*:

$$Ru(t) + \int_0^T K(t, \sigma) u(\sigma)\, d\sigma = -F^*(T-t) Q_T f_{(x_0, \phi, \psi)}(T) - \int_t^T F^*(\sigma - t) Q f_{(x_0, \phi, \psi)}(\sigma)\, d\sigma, \qquad t \in [0, T], \tag{2.11}$$

where the kernel operator function is defined by

$$K(t, \sigma) = F^*(T-t) Q_T F(T-\sigma) + \int_t^T F^*(s-t) Q F(s-\sigma)\, ds, \qquad (t, \sigma) \in [0, T]^2. \tag{2.12}$$

Here we also use the convention that (*see* (1.32) *and* (1.28))

$$F(t) \equiv 0 \quad \textit{whenever } t < 0.$$

Proof. Since we can deduce that

$$(\Gamma^* Q \Gamma + \Gamma_T^* Q_T \Gamma_T) u = \int_0^T K(\cdot, \sigma) u(\sigma)\, d\sigma,$$

this result is a direct consequence of the optimal control relation (2.8). The uniqueness can be proved by the fact that $\Phi > 0$. □

COROLLARY 2. *The optimum of $J(u)$ on $\mathscr{U}$ is a quadratic form with respect to the initial condition $(x_0, \phi, \psi) \in Z$ and is given by*

$$J^*(x_0, \phi, \psi) = \min_{u \in \mathscr{U}} J(u; x_0, \phi, \psi) = \left\langle P \begin{pmatrix} x_0 \\ \phi \\ \psi \end{pmatrix}, \begin{pmatrix} x_0 \\ \phi \\ \psi \end{pmatrix} \right\rangle_Z, \tag{2.13}$$

where $P \in \mathscr{L}(Z)$ is a nonnegative operator given by

$$P = \tilde{f}^* Q \tilde{f} + \tilde{f}_T^* Q_T \tilde{f}_T - (\Gamma^* Q \tilde{f} + \Gamma_T^* Q_T \tilde{f}_T)^* \Phi^{-1} (\Gamma^* Q \tilde{f} + \Gamma_T^* Q_T \tilde{f}_T), \tag{2.14}$$

where $\tilde{f} \in \mathscr{L}(Z; \mathscr{X})$ and $\tilde{f}_T \in \mathscr{L}(Z; X)$ are defined by

$$\tilde{f} \begin{pmatrix} x_0 \\ \phi \\ \psi \end{pmatrix} = f_{(x_0, \phi, \psi)}(\cdot) \quad \textit{and} \quad \tilde{f}_T \begin{pmatrix} x_0 \\ \phi \\ \psi \end{pmatrix} = f_{(x_0, \phi, \psi)}(T),$$

respectively.

The proof is simply a completion of squares and hence is omitted here.

3. Semicausal optimality principle. To implement the optimal control $u(t)$ by the attainable information involving only the past state $\{x(\sigma): 0 \leqq \sigma \leqq t\}$, we must try to transfer the effect of the future state segment $\{x(\sigma): t < \alpha \leqq T\}$ to the real time optimal control $u(t)$ by seeking appropriate feedback. Similarly, as we have done in [18], there are two important notions to be introduced here.

DEFINITION 1. Define a truncation operator P_ξ by

$$(P_\xi u)(t) = \begin{cases} u(t), & 0 \leqq t \leqq \xi, \\ 0, & \xi < t \leqq T, \end{cases} \tag{3.1}$$

where $0 \leqq \xi \leqq T$ is arbitrarily given.

Then, both P_ξ and $(I - P_\xi) \in \mathscr{L}(\mathscr{U})$ are projections and it is obvious that

$$\begin{aligned} &P_\xi R = RP_\xi, \quad (I - P_\xi)R = R(I - P_\xi), \\ &R^{-1}P_\xi = P_\xi R^{-1}, \quad R^{-1}(I - P_\xi) = (I - P_\xi)R^{-1}. \end{aligned} \tag{3.2}$$

Denote

$$\mathscr{U}_\xi = (I - P_\xi)\mathscr{U} \subset \mathscr{U}.$$

DEFINITION 2. For a given (x_0, ϕ, ψ) and admissible control $u(\cdot)$, define a semicausal trajectory $x_\xi(\cdot)$ with parameter ξ by

$$x_\xi(t) = f_{(x_0,\phi,\psi)}(t) + \int_0^t F(t-\sigma)P_\xi u(\sigma)\, d\sigma, \qquad t \in [0, T]. \tag{3.3}$$

According to these definitions, the following relations are valid:

$$\begin{aligned} &u = P_\xi u + (I - P_\xi)u, \qquad x_\xi = f_{(x_0,\phi,\psi)} + \Gamma P_\xi u, \\ &x = x_\xi + \Gamma(I - P_\xi)u, \qquad x(T) = x_\xi(T) + \Gamma_T(I - P_\xi)u, \end{aligned} \tag{3.4}$$

where $0 \leqq \xi \leqq T$.

Define

$$\Phi_\xi^+ = (I - P_\xi)\Phi \,|\, \mathscr{U}_\xi, \tag{3.5}$$

where $0 \leqq \xi \leqq T$ is a parameter.

Similarly to [18, Lemma 4], $\Phi_\xi^+ \in \mathscr{L}(\mathscr{U}_\xi)$ is self-adjoint and coercively positive, so that Φ_ξ^+ is invertible and, moreover,

$$\|(\Phi_\xi^+)^{-1}\|_{\mathscr{L}(\mathscr{U}_\xi)} \leqq \text{const.} \tag{3.6}$$

where the constant is uniform for $0 \leqq \xi \leqq T$.

Now define another operator $N_\xi \in \mathscr{L}(\mathscr{X} \times X)$ by

$$N_\xi = I_{\mathscr{X}\times X} - \begin{pmatrix} \Gamma \\ \Gamma_T \end{pmatrix} (\Phi_\xi^+)^{-1}(I - P_\xi)(\Gamma^* Q, \Gamma_T^* Q_T). \tag{3.7}$$

LEMMA 6. *Let $\xi \in [0, T]$ be arbitrarily given. The optimal state trajectory $x(\cdot)$ and the corresponding semicausal trajectory $x_\xi(\cdot)$ are related by*

$$\begin{pmatrix} x(\cdot) \\ x(T) \end{pmatrix} = N_\xi \begin{pmatrix} x_\xi(\cdot) \\ x_\xi(T) \end{pmatrix}. \tag{3.8}$$

Proof. For any given initial data $\{x_0, \phi, \psi\}$, let $\{u(\cdot), x(\cdot)\}$ be the optimal control process. From (2.9) and (3.4) it follows that

$$Ru + (\Gamma^* Q \Gamma + \Gamma_T^* Q_T \Gamma_T)(I - P_\xi)u = -(\Gamma^* Q x_\xi + \Gamma_T^* Q_T x_\xi(T)). \tag{3.9}$$

Apply the operator $(I - P_\xi)$ on both sides of (3.9) and use the invertibility of the operator Φ_ξ^+; then

$$(I - P_\xi)u = -(\Phi_\xi^+)^{-1}(I - P_\xi)(\Gamma^* Q, \Gamma_T^* Q_T) \begin{pmatrix} x_\xi(\cdot) \\ x_\xi(T) \end{pmatrix}. \tag{3.10}$$

Substitute (3.10) into the expressions for $x(\cdot)$ and $x(T)$ in (3.4); then (3.8) is obtained. □

Remark 1. If $\xi = 0$, then $x_\xi(t) = f_{(x_0,\phi,\psi)}(t)$, for any $t \in (0, T]$. We have

$$N_{\xi=0} = I_{\mathscr{X}\times X} - \begin{pmatrix} \Gamma \\ \Gamma_T \end{pmatrix} \Phi^{-1}(\Gamma^* Q, \Gamma_T^* Q_T),$$

since $(\Phi_{\xi=0}^+)^{-1} = \Phi^{-1}$ and $I - P_0 = I$. Hence it follows that when $\xi = 0$,

$$N_\xi \begin{pmatrix} x_\xi(\cdot) \\ x_\xi(T) \end{pmatrix} = N_0 \begin{pmatrix} f_{(x_0,\phi,\psi)}(\cdot) \\ f_{(x_0,\phi,\psi)}(T) \end{pmatrix}$$

$$= \begin{pmatrix} f_{(x_0,\phi,\psi)}(\cdot) - \Gamma\Phi^{-1}(\Gamma^* Q f_{(x_0,\phi,\psi)} + \Gamma_T^* Q_T f_{(x_0,\phi,\psi)}(T)) \\ f_{(x_0,\phi,\psi)}(T) - \Gamma_T\Phi^{-1}(\Gamma^* Q f_{(x_0,\phi,\psi)} + \Gamma_T^* Q_T f_{(x_0,\phi,\psi)}(T)) \end{pmatrix}$$

(by the open-loop relation (2.8))

$$= \begin{pmatrix} f_{(x_0,\phi,\psi)}(\cdot) + \Gamma\Phi^{-1}(\Phi u) \\ f_{(x_0,\phi,\psi)}(T) + \Gamma_T\Phi^{-1}(\Phi u) \end{pmatrix}$$

$$= \begin{pmatrix} f_{(x_0,\phi,\psi)}(\cdot) + \Gamma u \\ f_{(x_0,\phi,\psi)}(T) + \Gamma_T u \end{pmatrix} = \begin{pmatrix} x(\cdot) \\ x(T) \end{pmatrix},$$

where u is the optimal control corresponding to the given initial data (x_0, ϕ, ψ), and $x(\cdot)$ is the optimal trajectory corresponding to u.

Therefore, even if $\xi = 0$, Lemma 6 and (3.8) still hold. The reason is that for the optimal process, the optimal control $u(\cdot)$ is uniquely determined by the initial data (x_0, ϕ, ψ), so the optimal trajectory $x(\cdot)$ and its terminal value $x(T)$, in turn, are also uniquely determined by the initial data (x_0, ϕ, ψ) only.

THEOREM 3 (Optimality Principle). *$u(\cdot)$ is an optimal control if and only if*

$$u(t) = -R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T) N_t \begin{pmatrix} x_t(\cdot) \\ x_t(T) \end{pmatrix}, \qquad t \in [0, T], \tag{3.11}$$

where $x_t(\cdot)$ is the corresponding semicausal trajectory with parameter t, N_t is given by (3.7), *and the operators $\Gamma^*(t) \in \mathscr{L}(\mathscr{X}; U)$ and $\Gamma_T^*(t) \in \mathscr{L}(X; U)$ are defined by*

$$\begin{aligned} &\Gamma^*(t)y = \int_t^T F^*(\sigma - t)y(\sigma)\, d\sigma, \quad \forall y \in \mathscr{X}, \\ &\Gamma_T^*(t)y_1 = F^*(T - t)y_1, \quad \forall y_1 \in X, \end{aligned} \tag{3.12}$$

respectively.

Proof. (i) If $u(\cdot)$ is an optimal control, from (2.9) and (3.8) it follows that

$$u = -R^{-1}(\Gamma^* Q, \Gamma_T^* Q_T) N_\xi \begin{pmatrix} x_\xi \\ x_\xi(T) \end{pmatrix}, \tag{3.13}$$

where $\xi \in [0, T]$ is arbitrary. Note (2.10) and (3.12), and take $\xi = t$ for the real time value $u(t)$; then (3.13) reduces to (3.11).

(ii) Conversely, to show that if $u(\cdot)$ satisfies (3.11) then it must be the optimal control, it is enough to show that the control $u(\cdot)$, which satisfies (3.11), is unique. This amounts to showing that $f_{(x_0,\phi,\psi)} = 0$ and $u(\cdot)$ satisfies (3.11) only if $u(t) \equiv 0$ on $[0, T]$.

Indeed, since $f_{(x_0,\phi,\psi)} = 0$, we have $x_t = \Gamma P_t u$. By (3.11), it can be estimated that

$$\|u(t)\| \leqq R^{-1}\| \|(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T)\|_{\mathscr{L}(\mathscr{X}\times X; U)} \|N_t\|_{\mathscr{L}(\mathscr{X}\times X)} \left\| \begin{pmatrix} x_t \\ x_t(T) \end{pmatrix} \right\|_{\mathscr{X}\times X} \tag{3.14}$$

Since (1.32) implies $F(t)$ is piecewise strongly continuous and uniformly bounded on the finite interval $[0, T]$, from (3.12), (2.1), (2.2), and (3.7), it follows that

$$\begin{aligned} &\|(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T)\|_{\mathscr{L}(\mathscr{X}\times X;U)} \leqq \text{const.},\\ &\|N_t\|_{\mathscr{L}(\mathscr{X}\times X)} \leqq \text{const.}, \end{aligned} \tag{3.15}$$

and

$$\begin{aligned} \left\|\begin{pmatrix} x_t \\ x_t(T)\end{pmatrix}\right\|_{\mathscr{X}\times X} &= \left\|\begin{pmatrix} \Gamma P_t u \\ \Gamma_T P_t u\end{pmatrix}\right\|_{\mathscr{X}\times X} \leqq \text{const.}\,(\|\Gamma\|_{\mathscr{L}(\mathscr{U};\mathscr{X})} + \|\Gamma_T\|_{\mathscr{L}(\mathscr{U};\mathscr{X})})\|P_t u\|_{\mathscr{U}} \\ &= \text{const.}\,\|P_t u\|_{\mathscr{U}}, \end{aligned} \tag{3.16}$$

where all the constants are independent of $t \in [0, T]$. Finally, since

$$\|P_t u\|_{\mathscr{U}} = \left\{\int_0^t \|u(s)\|^2\, ds\right\}^{1/2}, \qquad t \in [0, T], \tag{3.17}$$

all these estimations imply that

$$\|u(t)\|^2 \leqq \text{const.} \int_0^t \|u(s)\|^2\, ds, \qquad t \in [0, T], \tag{3.18}$$

which allows us to use the Gronwall–Bellman inequality to obtain that $u(t) \equiv 0$. □

This optimality principle has qualitatively demonstrated the causal dependence of the optimal control $u(t)$ on the past information $x_t(\cdot)$ or $P_t u$.

4. Fredholm synthesis equation. Let $\xi \in [0, T]$. Define an operator function

$$M_\xi(t, \sigma) = -R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T) N_\xi \begin{pmatrix} F(\cdot - \sigma) \\ F(T - \sigma)\end{pmatrix}, \tag{4.1}$$

where $(t, \sigma) \in [0, T]\times[0, T]$, and $M_\xi(t, \sigma) \in \mathscr{L}(U)$.

LEMMA 7. *The following identities hold*:

$$(\Phi_\xi^+)^{-1}(I - P_\xi)(\Gamma^* Q, \Gamma_T^* Q_T) = (I - P_\xi) R^{-1}(\Gamma^* Q, \Gamma_T^* Q_T) N_\xi : \mathscr{X}\times X \to \mathscr{U}_\xi, \tag{4.2}$$

and

$$\begin{pmatrix} \Gamma \\ \Gamma_T\end{pmatrix} (\Phi_\xi^+)^{-1}(I - P_\xi) = N_\xi \begin{pmatrix} \Gamma \\ \Gamma_T\end{pmatrix} (I - P_\xi) R^{-1} : \mathscr{U} \to \mathscr{X}\times X. \tag{4.3}$$

Proof. The proof is simply a verification with the use of definitions, the idempotent property of $(I - P_\xi)$, and (3.2). Since it is similar to Lemma 6 of [8], it is omitted here.

THEOREM 4 (Synthesis Equation). *For any given $\xi \in [0, T]$, there exists a unique solution $M_\xi(t, \sigma)$ of the following equation*:

$$M_\xi(t, \sigma) + \int_\xi^T R^{-1} K(t, s) M_\xi(s, \sigma)\, ds = -R^{-1} K(t, \sigma), \qquad (t, \sigma) \in [0, T]\times[0, T], \tag{4.4}$$

where $K(t, \sigma)$ is given by (2.12). *The solution is given by* (4.1), *which is strongly continuous in t and piecewise continuous in σ and such that*

$$\|M_\xi(t, \sigma)\| \leqq \text{const.}, \qquad \forall (\xi, t, \sigma) \in [0, T]^3. \tag{4.5}$$

Proof. First we show that $M_\xi(t, \sigma)$ given by (4.1) is a solution of the equation (4.4). Besides, it is easy to see that this $M_\xi(t, \sigma)$ does satisfy the required conditions

of continuity and boundedness from the definition. We check that (4.4) is satisfied by (4.1) as follows:

$$M_\xi(t,\sigma) = -R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T)N_\xi \begin{pmatrix} F(\cdot-\sigma) \\ F(T-\sigma) \end{pmatrix}$$

$$= -R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T)\left\{ I_{\mathscr{X}\times X} - \begin{pmatrix} \Gamma \\ \Gamma_T \end{pmatrix} (\Phi_\xi^+)^{-1}(I-P_\xi)(\Gamma^*Q, \Gamma_T^*Q_T) \right\}$$

$$\cdot \begin{pmatrix} F(\cdot-\sigma) \\ F(T-\sigma) \end{pmatrix}$$

$$= -R^{-1}(\Gamma^*(t)QF(\cdot-\sigma) + \Gamma_T^*(t)Q_TF(T-\sigma))$$

$$+ R^{-1}(\Gamma^*(t)Q\Gamma + \Gamma_T^*(t)Q_T\Gamma_T)(\Phi_\xi^+)^{-1}(I-P_\xi)(\Gamma^*Q, \Gamma_T^*Q_T)\begin{pmatrix} F(\cdot-\sigma) \\ F(T-\sigma) \end{pmatrix}$$

$$= -R^{-1}(\Gamma^*(t)QF(\cdot-\sigma) + \Gamma_T^*(t)Q_TF(T-\sigma))$$

(by (4.2))

$$+ R^{-1}(\Gamma^*(t)Q\Gamma + \Gamma_T^*(t)Q_T\Gamma_T)(I-P_\xi)R^{-1}(\Gamma^*Q, \Gamma_T^*Q_T)N_\xi \begin{pmatrix} F(\cdot-\sigma) \\ F(T-\sigma) \end{pmatrix}$$

$$= -R^{-1}(\Gamma^*(t)QF(\cdot-\sigma) + \Gamma_T^*(t)Q_TF(T-\sigma))$$

(by (4.1))

$$-R^{-1}(\Gamma^*(t)Q\Gamma + \Gamma_T^*(t)Q_T\Gamma_T)(I-P_\xi)M_\xi(\cdot,\sigma). \tag{4.6}$$

From (3.12), (2.1), (2.2), and (2.12), it follows that

$$R^{-1}(\Gamma^*(t)QF(\cdot-\sigma) + \Gamma_T^*(t)Q_TF(T-\sigma)) = R^{-1}K(t,\sigma), \tag{4.7}$$

and

$$\begin{aligned} &R^{-1}(\Gamma^*(t)Q\Gamma + \Gamma_T^*(t)Q_T\Gamma_T)(I-P_\xi)M_\xi(\cdot,\sigma) \\ &\quad = \int_0^T R^{-1}K(t,s)(I-P_\xi)M_\xi(s,\sigma)\,ds = \int_\xi^T R^{-1}K(t,s)M_\xi(s,\sigma)\,ds. \end{aligned} \tag{4.8}$$

Substitute (4.7) and (4.8) into (4.6); then we obtain

$$M_\xi(t,\sigma) = -R^{-1}K(t,\sigma) - \int_\xi^T R^{-1}K(t,s)M_\xi(s,\sigma)\,ds.$$

This last relation shows that the $M_\xi(t,\sigma)$ given by (4.1) is exactly a solution of the equation (4.4).

Next we prove the uniqueness. This can be done by showing that the corresponding homogeneous equation

$$\tilde{M}_\xi(t,\sigma) + \int_\xi^T R^{-1}K(t,s)\tilde{M}_\xi(s,\sigma)\,ds = 0, \qquad (t,\sigma)\in[0,T]^2 \tag{4.9}$$

admits only the null solution $\tilde{M}_\xi(t,\sigma)\equiv 0$. Indeed, (4.9) restricted on $(t,\sigma)\in[\xi,T]\times[0,T]$ can be rewritten as

$$\Phi_\xi^+\tilde{M}(\cdot,\sigma) = 0 \quad \text{for each } \sigma\in[0,T]. \tag{4.10}$$

Since Φ_ξ^+ is invertible in $\mathscr{L}(\mathscr{U}_\xi)$, we have

$$\tilde{M}_\xi(s,\sigma)\equiv 0 \quad \text{for } s\in[\xi, T] \text{ and } \sigma\in[0, T]. \tag{4.11}$$

Then substitute (4.11) into the integral term of (4.9); we finally obtain

$$\tilde{M}_\xi(t,\sigma)\equiv 0 \quad \text{for } t\in[0, T] \text{ and } \sigma\in[0, T]. \tag{4.12}$$

Thus the proof is completed. □

This result means that the inversion $(\Phi_\xi^+)^{-1}$ has been transferred into solving a linear integral equation (4.4) of Fredholm type. Thus it is possible to obtain the following feedback optimal control by using the semicausal trajectory.

THEOREM 5. *$u(\cdot)$ is the optimal control if and only if*

$$\begin{aligned} u(t) &= -R^{-1}[\Gamma^*(t)Qx_t+\Gamma_T^*(t)Q_Tx_t(T)] \\ &\quad -\int_t^T M_t(t,\sigma)R^{-1}[\Gamma^*(\sigma)Qx_t+\Gamma_T^*(\sigma)Q_Tx_t(T)]\,d\sigma, \qquad t\in[0, T], \end{aligned} \tag{4.13}$$

where $M_\xi(t,\sigma)$ is the unique solution of the linear integral equation (4.4) *of Fredholm type, and $x_t(\cdot)$ is the corresponding semicausal trajectory with real time t as parameter value.*

Proof. If $u(\cdot)$ is the optimal control, by the previous results (3.11), (3.7), and (4.3), it follows that

$$\begin{aligned} u(t) &= -R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T)N_t\begin{pmatrix} x_t \\ x_t(T)\end{pmatrix} \\ &= -R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T)\begin{pmatrix} x_t \\ x_t(T)\end{pmatrix} + R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T)\begin{pmatrix}\Gamma \\ \Gamma_T\end{pmatrix} \\ &\qquad \cdot(\Phi_t^+)^{-1}(I-P_t)(\Gamma^*Q, \Gamma_T^*Q_T)\begin{pmatrix} x_t \\ x_t(T)\end{pmatrix} \\ &= -R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_t)\begin{pmatrix} x_t \\ x_t(T)\end{pmatrix} + R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T)N_t\begin{pmatrix}\Gamma \\ \Gamma_T\end{pmatrix} \\ &\qquad \cdot(I-P_t)R^{-1}(\Gamma^*Q, \Gamma_T^*Q_T)\begin{pmatrix} x_t \\ x_t(T)\end{pmatrix} \\ &= -R^{-1}[\Gamma^*(t)Qx_t+\Gamma_T^*(t)Q_Tx_t(T)] \\ &\quad + R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T)N_t \\ &\qquad \cdot\begin{pmatrix} \int_t^T F(\cdot-\sigma)R^{-1}[\Gamma^*(\sigma)Qx_t+\Gamma_T^*(\sigma)Q_Tx_t(T)]\,d\sigma \\ \int_t^T F(T-\sigma)R^{-1}[\Gamma^*(\sigma)Qx_t+\Gamma_T^*(\sigma)Q_Tx_t(T)]\,d\sigma\end{pmatrix} \\ &= -R^{-1}[\Gamma^*(t)Qxt+\Gamma_T^*(t)Q_Tx_t(T)] \\ &\qquad + \int_t^T R^{-1}(\Gamma^*(t)Q, \Gamma_T^*(t)Q_T)N_t\begin{pmatrix} F(\cdot-\sigma) \\ F(T-\sigma)\end{pmatrix} \\ &\qquad \cdot R^{-1}[\Gamma^*(\sigma)Qx_t+\Gamma_T^*(\sigma)Q_Tx_t(T)]\,d\sigma \\ &= -R^{-1}[\Gamma^*(t)Qx_t+\Gamma_T^*(t)Q_Tx_t(T)] \\ &\qquad -\int_t^T M_t(t,\sigma)R^{-1}[\Gamma^*(\sigma)Qx_t+\Gamma_T^*(\sigma)Q_Tx_t(T)]\,d\sigma, \end{aligned} \tag{4.14}$$

where we have made use of the convention (1.28). Equation (4.14) shows that the optimal control actually satisfies (4.13).

Conversely, if the initial data (x_0, ϕ, ψ) is given, then there exists only one control $u(\cdot)$, which satisfies the relation (4.13). This can be shown in a similar way by the Gronwall-Bellman inequality, as in the proof of Theorem 4. Thus it is concluded that (4.13) is necessary and sufficient for $u(\cdot)$ to be optimal. □

THEOREM 6 (semicausal trajectory feedback). *$u(\cdot)$ is the optimal control if and only if*

$$u(t) = -\pi(T, t)Q_T x_t(T) - \int_t^T \pi(s, t)Qx_t(s)\,ds, \qquad t \in [0, T], \tag{4.15}$$

where

$$\pi(s, t) = R^{-1}F^*(s-t) + \int_t^s M_t(t, \sigma)R^{-1}F^*(s-\sigma)\,d\sigma, \tag{4.16}$$

and $M_\xi(t, \sigma)$ and $x_t(\cdot)$ are described as in Theorem 5.

Proof. Substitute (3.12) into (4.13), to obtain

$$\begin{aligned} u(t) &= -R^{-1}\left[F^*(T-t)Q_T x_t(T) + \int_t^T F^*(s-t)Qx_t(s)\,ds\right] \\ &\quad - \int_t^T M_t(t, \sigma)R^{-1}\left[F^*(T-\sigma)Q_T x_t(T) + \int_\sigma^T F^*(s-\sigma)Qx_t(s)\,ds\right]d\sigma \\ &= -\left[R^{-1}F^*(T-t) + \int_t^T M_t(t, \sigma)R^{-1}F^*(T-\sigma)\,d\sigma\right]Q_T x_t(T) \\ &\quad - \int_t^T \left[R^{-1}F^*(s-t) + \int_t^s M_t(t, \sigma)R^{-1}F^*(s-\sigma)\,d\sigma\right]Qx_t(s)\,ds \\ &= -\pi(T, t)Q_T x_t(T) - \int_t^T \pi(s, t)Qx_t(s)\,ds, \qquad t \in [0, T]. \end{aligned}$$

Conversely, (4.13) can be deduced from (4.15) so these two relations are equivalent. Thus the conclusion is true. □

Remark 2. Since the semicausal state trajectory $x_t(\cdot)$ with real time t as a running parameter is determined only by the past control $\{u(\sigma): \sigma \leqq t\}$, and the feedback operator function $\pi(s, t)$ is known provided the solution $M_\xi(t, \sigma)$ of the equation (4.4) is available by any means (analytically or numerically), the obtained result (4.15) is justified to be a closed-loop solution of the described optimal control problem.

5. Closed-loop optimal control. In this section, we present an expression for the real time optimal control $u(t)$ by state feedback that involves $\{x(\sigma): \sigma \leqq t\}$, based on the semicausal trajectory feedback given by Theorem 6.

A clarification of the relationship between the state function $x(t)$ and the corresponding semicausal trajectory $x_\xi(t)$ will pave the way to the closed-loop synthesis of the optimal control.

LEMMA 8. *Let $G(t)$ be the fundamental solution associated with* (1.6) *and let $F(t)$ be defined by* (1.32). *Then, the following relation holds*:

$$F(t+s) = G(t)F(s) + \sum_{i=1}^{n} \int_0^{\min(t, h_i)} G(t-\sigma)A_i F(s+\sigma-h_i)\,d\sigma \quad \text{for } t \geqq 0,\ s \geqq 0. \tag{5.1}$$

Proof. The equality can be verified by the use of (1.23) and (1.32). □

LEMMA 9. *Let $x(\cdot)$ be any state function with an initial data $(x_0, \phi, \psi) \in Z$ and an admissible control $u \in \mathcal{U}$, and let $x_\xi(\cdot)$ be the corresponding semicausal trajectory with parameter $\xi \in [0, T]$. Then, the following relation holds:*

$$x_\xi(t) = \begin{cases} x(t) & \text{for } t \leqq \xi, \\ G(t-\xi)x(\xi) + \sum_{i=1}^{n} \int_{\xi}^{\min(t,\xi+h_i)} G(t-\eta)A_i x(\eta - h_i)\, d\eta \end{cases}$$

$$+ \sum_{i=1}^{n} \int_{\min(\xi,h_i)}^{h_i} G(t-\eta)A_i \phi(\eta - h_i)\, d\eta \tag{5.2}$$

$$+ \sum_{j=1}^{m} \int_{\min(\xi,r_j)}^{r_j} G(t-\eta)B_j \psi(\eta - r_j)\, d\eta \quad \text{for } t > \xi.$$

Proof. It is enough to prove (5.2) only for $t > \xi$. In fact, for $t > \xi$,

$$x_\xi(t) \underset{(3.3)}{=} f_{(x_0,\phi,\psi)}(t) + \int_0^\xi F(t-\sigma)u(\sigma)\, d\sigma$$

$$\underset{(1.31)}{=} G(t)x_0 + \int_0^t G(t-\sigma)\left\{ \sum_{i=1}^{n} A_i \phi(\sigma - h_i) + \sum_{j=1}^{m} B_j \psi(\sigma - r_j) \right\} d\sigma$$

$$+ \int_0^\xi F(t-\sigma)u(\sigma)\, d\sigma$$

$$\underset{(5.1)}{\overset{(1.23)}{=}} G(t-\xi)G(\xi)x_0 + \sum_{i=1}^{n} \int_{\xi}^{\min(t-\xi,h_i)} G(t-\xi-s)A_i G(\xi+s-h_i)\, ds\, x_0$$

$$+ G(t-\xi) \int_0^\xi G(\xi-\sigma)\left\{ \sum_{i=1}^{n} A_i \phi(\sigma - h_i) + \sum_{j=1}^{m} B_j \psi(\sigma - r_j) \right\} d\sigma$$

$$+ \sum_{i=1}^{n} \int_0^\xi \int_0^{\min(t-\xi,h_i)} G(t-\xi-s)A_i G(\xi-\sigma+s-h_i)\, ds$$

$$\cdot \left\{ \sum_{i=1}^{n} A_i \phi(\sigma - h_i) + \sum_{j=1}^{m} B_j \psi(\sigma - r_j) \right\} d\sigma$$

$$+ G(t-\xi) \int_0^\xi F(\xi-\sigma)u(\sigma)\, d\sigma$$

$$+ \sum_{i=1}^{n} \int_0^\xi \int_0^{\min(t-\xi,h_i)} G(t-\xi-s)A_i F(\xi-\sigma+s-h_i)\, ds\, u(\sigma)\, d\sigma$$

$$+ \int_\xi^t G(t-\sigma)\left\{ \sum_{i=1}^{n} A_i \phi(\sigma - h_i) + \sum_{j=1}^{m} B_j \psi(\sigma - r_j) \right\} d\sigma$$

$$= G(t-\xi)\left\{ G(\xi)x_0 + \int_0^\xi G(\xi-\sigma)\left\{ \sum_{i=1}^{n} A_i \phi(\sigma - h_i) + \sum_{j=1}^{m} B_j \psi(\sigma - r_j) \right\} d\sigma \right.$$

$$\left. + \int_0^\xi F(\xi-\sigma)u(\sigma)\, d\sigma \right\}$$

$$+\sum_{i=1}^{n}\int_{\xi}^{\min(t,\xi+h_i)} G(t-\eta)A_i\Bigg[G(\eta-h_i)x_0+\int_0^{\xi} G(\eta-\sigma-h_i)\cdot\Bigg\{\sum_{i=1}^{n}A_i\phi(\sigma-h_i)+\sum_{j=1}^{m}B_j\psi(\sigma-r_j)\Bigg\}d\sigma+\int_0^{\xi}F(\eta-\sigma-h_i)u(\sigma)\,d\sigma\Bigg]d\eta$$

$$+\int_{\xi}^{t}G(t-\sigma)\Bigg\{\sum_{i=1}^{n}A_i\phi(\sigma-h_i)+\sum_{j=1}^{m}B_j\psi(\sigma-r_j)\Bigg\}d\sigma$$

(note that $G(t)=0$ and $F(t)=0$ whenever $t<0$)

$$=G(t-\xi)x(\xi)+\sum_{i=1}^{n}\int_{\xi}^{\min(t,\xi+h_i)}G(t-\eta)A_ix(\eta-h_i)\,d\eta+\sum_{i=1}^{n}\int_{\min(\xi,h_i)}^{h_i}G(t-\eta)A_i\phi(\eta-h_i)\,d\eta+\sum_{j=1}^{m}\int_{\min(\xi,r_j)}^{r_j}G(t-\eta)B_j\psi(\eta-r_j)\,d\eta.$$

Therefore, (5.2) is valid. □

Based on these above results, the closed-loop optimal control is achieved as follows.

THEOREM 7 (closed-loop optimal control). *$u(\cdot)$ is the optimal control if and only if it is given by the following linear state feedback*:

$$u(t)=-\Bigg[H(t,t)x(t)+\sum_{i=1}^{n}\int_{t}^{\min(T,t+h_i)}H(t,\eta)A_ix(\eta-h_i)\,d\eta+\sum_{i=1}^{n}\int_{\min(t,h_i)}^{h_i}H(t,\eta)A_ix(\eta-h_i)\,d\eta+\sum_{j=1}^{m}\int_{\min(t,r_j)}^{r_j}H(t,\eta)B_ju(\eta-r_j)\,d\eta\Bigg],\qquad t\in[0,T],\tag{5.3}$$

where $x(\cdot)$ is the corresponding state function, and

(5.4)

$$H(t,\eta)=\pi(T,t)Q_TG(T-\eta)+\int_t^T\pi(s,t)QG(s-\eta)\,ds,\qquad(t,\eta)\in[0,T]\times[0,T]$$

and $\pi(s,t)$ is given by (4.16).

Proof. Substitute (5.2) into (4.15), then the optimal control $u(t)$ is given by

$$u(t)=-\pi(T,t)Q_Tx_t(T)-\int_t^T\pi(s,t)Qx_t(s)\,ds$$

$$=-\pi(T,t)Q_T\Bigg\{G(T-t)x(t)+\sum_{i=1}^{n}\int_t^{\min(T,t+h_i)}G(T-\eta)A_ix(\eta-h_i)\,d\eta+\sum_{i=1}^{n}\int_{\min(t,h_i)}^{h_i}G(T-\eta)A_i\phi(\eta-h_i)\,d\eta+\sum_{j=1}^{m}\int_{\min(t,r_j)}^{r_j}G(T-\eta)B_j\psi(\eta-r_j)\,d\eta\Bigg\}$$

$$-\int_t^T\pi(s,t)Q\Bigg\{G(s-t)x(t)+\sum_{i=1}^{n}\int_t^{\min(s,t+h_i)}G(s-\eta)A_ix(\eta-h_i)\,d\eta$$

$$+\sum_{i=1}^{n}\int_{\min(t,h_i)}^{h_i} G(s-\eta)A_i\phi(\eta-h_i)\,d\eta$$

$$+\sum_{j=1}^{m}\int_{\min(t,r_j)}^{r_j} G(s-\eta)B_j\psi(\eta-r_j)\,d\eta\Bigg\}\,ds$$

$$=-\Bigg[\pi(T,t)Q_TG(T-t)+\int_t^T \pi(s,t)QG(s-t)\,ds\Bigg]x(t)$$

$$-\sum_{i=1}^{n}\int_t^{\min(T,t+h_i)}\Bigg[\pi(T,t)Q_TG(T-\eta)+\int_t^T \pi(s,t)QG(s-\eta)\,ds\Bigg]$$

$$\cdot A_ix(\eta-h_i)\,d\eta$$

$$-\sum_{i=1}^{n}\int_{\min(t,h_i)}^{h_i}\Bigg[\pi(T,t)Q_TG(T-\eta)+\int_t^T \pi(s,t)QG(s-\eta)\,ds\Bigg]$$

$$\cdot A_i\phi(\eta-h_i)\,d\eta$$

$$-\sum_{j=1}^{m}\int_{\min(t,r_j)}^{r_j}\Bigg[\pi(T,t)Q_TG(T-\eta)+\int_t^T \pi(s,t)QG(s-\eta)\,ds\Bigg]$$

$$\cdot B_j\psi(\eta-r_j)\,d\eta$$

$$=-\Bigg[H(t,t)x(t)+\sum_{i=1}^{n}\int_t^{\min(T,t+h_i)} H(t,\eta)A_ix(\eta-h_i)\,d\eta$$

$$+\sum_{i=1}^{n}\int_{\min(t,h_i)}^{h_i} H(t,\eta)A_ix(\eta-h_i)\,d\eta$$

$$+\sum_{j=1}^{m}\int_{\min(t,r_j)}^{r_j} H(t,\eta)B_ju(\eta-r_j)\,d\eta\Bigg],\qquad t\in[0,T],$$

where $\phi(\eta-h_i)=x(\eta-h_i)$ for $\min(t,h_i)\leqq\eta<h_i$ so that $-h_i\leqq\eta-h_i<0$, $i=1,\cdots,n$, and $\psi(\eta-r_j)=u(\eta-r_j)$ for $\min(t,r_j)\leqq\eta<r_j$ so that $-r_j\leqq\eta-r_j\leqq 0$, $j=1,\cdots,m$. This indicates that the optimal control $u(\cdot)$ satisfies the feedback relation (5.3). Conversely, if $u(\cdot)$ satisfies (5.3), we can deduce that $u(\cdot)$ satisfies (4.15). By Theorem 6, this $u(\cdot)$ must be the optimal control. □

COROLLARY 1. *For $t\in[\max(h_n,r_m),T]$, the optimal control is given by*

$$u(t)=-\Bigg[H(t,t)x(t)+\sum_{i=1}^{n}\int_t^{\min(T,t+h_i)} H(t,\eta)A_ix(\eta-h_i)\,d\eta\Bigg]. \tag{5.5}$$

Therefore the last two summation parts depending on the initial data ϕ and ψ have effects only on the small interval $[0,\max(h_n,r_m)]$.

COROLLARY 2. *If $B_1=\cdots=B_m=0$, i.e., there is no time delay in the control variable of the system* (1.1), *then the optimal control is given by*

$$u(t)=-\Bigg[L(t,t)x(t)+\sum_{i=1}^{n}\int_t^{\min(T,t+h_i)} L(t,\eta)A_ix(\eta-h_i)\,d\eta$$

$$+\sum_{i=1}^{n}\int_{\min(t,h_i)}^{h_i} L(t,\eta)A_ix(\eta-h_i)\,d\eta\Bigg],\qquad t\in[0,T], \tag{5.6}$$

where $L(t,\eta)$ is given by

$$L(t,\eta)=R^{-1}B^*\Psi(t,\eta)+\int_t^T M_t(t,\sigma)R^{-1}B^*\Psi(\sigma,\eta)\,d\sigma, \tag{5.7}$$

where

$$\Psi(t,\eta)=G^*(T-t)Q_TG(T-\eta)+\int_t^T G^*(s-t)QG(s-\eta)\,ds \tag{5.8}$$

and $M_\xi(t,\sigma)$ *is the unique solution of the following integral equation*:

$$M_\xi(t,\sigma)+\int_\xi^T R^{-1}B^*\Psi(t,s)BM_\xi(s,\sigma)\,ds=-R^{-1}B^*\Psi(t,\sigma)B, \tag{5.9}$$

$$(t,\sigma)\in[0,T]\times[0,T].$$

Proof. Let $F(t)=G(t)B$ be substituted into (2.12), (4.4), (4.16), and (5.4); then this result is a consequence of the main result described by Theorem 7. □

COROLLARY 3. *If* $A_1=\cdots=A_n=0$, *i.e., there is no time delay in the state variable of the system* (1.1), *then the optimal control is given by*

$$u(t)=-\left[H(t,t)x(t)+\sum_{j=1}^m\int_{\min(t,r_j)}^{r_j}H(t,\eta)B_ju(\eta-r_j)\,d\eta\right], \tag{5.10}$$

where $H(t,\eta)$ *is given by* (5.4) *and* $\pi(s,t)$ *is given by* (4.16), *with*

$$F(t)=e^{At}B+e^{A(t-r_1)}B_1+\cdots+e^{A(t-r_m)}B_m. \tag{5.11}$$

COROLLARY 4. *If* $A_i=0$ $(i=1,\cdots,n)$ *and* $B_j=0$ $(j=1,\cdots,m)$, *then the optimal control is given by*

$$u(t)=-H(t,t)x(t),\qquad t\in[0,T], \tag{5.12}$$

where

$$H(t,t)=\pi(T,t)Q_TG(T-t)+\int_t^T\pi(s,t)QG(s-t)\,ds \tag{5.13}$$

in which

$$G(t)=e^{At}\qquad (G^*(t)=e^{A^*t},\ t\geqq 0,\ \textit{is the dual semigroup}), \tag{5.14}$$

$$\pi(s,t)=R^{-1}B^*\,e^{A^*(s-t)}+\int_t^s M_t(t,\sigma)R^{-1}B^*\,e^{A^*(s-\sigma)}\,d\sigma, \tag{5.15}$$

$M_\xi(t,\sigma)$ *is the unique solution of the equation* (4.4), *in which*

$$K(t,\sigma)=B^*\left\{e^{A^*(T-t)}Q_Te^{A(T-\sigma)}+\int_t^T e^{A^*(s-t)}Qe^{A(s-\sigma)}\,ds\right\}B. \tag{5.16}$$

To show that in this nondelay case ($A_i=0$, $i=1,\cdots,n$, $B_j=0$, $j=1,\cdots,m$), the closed-loop optimal control given by (5.12) does coincide with the standard result by the state feedback with the solution of a Riccati equation, we need the following lemmas.

Later in this section, we denote $G(t)=e^{At}$, $t\geqq 0$, with the previous convention that $G(t)\equiv 0$ whenever $t<0$.

LEMMA 10. *Define an operator function* $W(t,\xi)$ *by*

$$W(t,\xi)=G(t-\xi)-\Gamma(t)(\Phi_\xi^+)^{-1}(I-P_\xi)(\Gamma^*Q,\Gamma_T^*Q_T)\begin{pmatrix}G(\cdot-\xi)\\ G(T-\xi)\end{pmatrix},\qquad 0<\xi\leqq t\leqq T, \tag{5.17}$$

and it is a convention that

$$W(t,\xi)=0\quad \textit{for } t<\xi,$$

where $\Gamma(t)\in\mathscr{L}(\mathscr{U};X)$ *is defined by* (*see* (2.1))

$$\Gamma(t)u=\int_0^t G(t-\sigma)Bu(\sigma)\,d\sigma.$$

Then,

(*i*) *For any given initial condition, the optimal state trajectory* $x(\cdot)$ *satisfies*

$$x(t) = W(t, \xi)x(\xi), \qquad 0 < \xi \leqq t \leqq T. \tag{5.18}$$

(ii) $\sup_{0<\xi\leqq t\leqq T} \| W(t, \xi)\|_{\mathscr{L}(X)} \leqq \text{const.}$

(iii) $W(t, \eta)W(\eta, \xi) = W(t, \xi)$, $0 < \xi \leqq \eta \leqq t \leqq T$.

(iv) $W(t, \xi)$ *is strongly continuous in* $t \in [\xi, T]$ *and in* $\xi \in (0, T]$, *respectively.*

Proof. (i) Since

$$x(t) \overset{(3.4)}{=} x_\xi(t) + \Gamma(t)(I - P_\xi)u, \qquad 0 < \xi \leqq t \leqq T,$$

$$(I - P_\xi)u \overset{(3.10)}{=} -(\Phi_\xi^+)^{-1}(I - P_\xi)(\Gamma^* Q x_\xi + \Gamma_T^* Q_T x_\xi(T)),$$

$$x_\xi(t) \overset{(5.2)}{=} G(t - \xi)x(\xi), \qquad 0 < \xi \leqq t \leqq T,$$

it follows that

$$x(t) = \left(G(t-\xi) - \Gamma(t)(\Phi_\xi^+)^{-1}(I - P_\xi)(\Gamma^* Q, \Gamma_T^* Q_T) \begin{pmatrix} G(\cdot - \xi) \\ G(T - \xi) \end{pmatrix} \right) x(\xi) = W(t, \xi)x(\xi).$$

(ii) By (3.6) and (5.17), $W(t, \xi)$ is uniformly bounded in the $\mathscr{L}(X)$ norm.

(iii) Let $0 < \xi \leqq \eta \leqq t \leqq T$; then it follows from (5.17) and (4.2) that

$$\begin{aligned} W(t, \eta)W(\eta, \xi) &= G(t-\xi) - G(t-\eta)\Gamma(\eta)(\Phi_\xi^+)^{-1}(I - P_\xi)(\Gamma^* Q, \Gamma_T^* Q_T) \begin{pmatrix} G(\cdot - \xi) \\ G(T - \xi) \end{pmatrix} \\ &\quad - \Gamma(t)(\Phi_\eta^+)^{-1}(I - P_\eta)(\Gamma^* Q, \Gamma_T^* Q_T) \begin{pmatrix} G(\cdot - \eta) \\ G(T - \eta) \end{pmatrix} G(\eta - \xi) \\ &\quad + \Gamma(t)(\Phi_\eta^+)^{-1}(I - P_\eta)(\Gamma^* Q, \Gamma_T^* Q_T) \begin{pmatrix} G(\cdot - \eta) \\ G(T - \eta) \end{pmatrix} \cdot \\ &\quad \cdot \Gamma(\eta)(\Phi_\xi^+)^{-1}(I - P_\xi)(\Gamma^* Q, \Gamma_T^* Q_T) \begin{pmatrix} G(\cdot - \xi) \\ G(T - \xi) \end{pmatrix} \\ &= G(t-\xi) - G(t-\eta)\Gamma(\eta)(I - P_\xi)R^{-1}(\Gamma^* Q, \Gamma_T^* Q_T)N_\xi \begin{pmatrix} G(\cdot - \xi) \\ G(T - \xi) \end{pmatrix} \\ &\quad - \Gamma(t)(I - P_\eta)R^{-1}(\Gamma^* Q, \Gamma_T^* Q_T) \\ &\quad \cdot \left\{ N_\eta - N_\eta \begin{pmatrix} G(\cdot - \eta) \\ G(T - \eta) \end{pmatrix} \Gamma(\eta) \right. \\ &\quad \left. \cdot (I - P_\xi)R^{-1}(\Gamma^* Q, \Gamma_T^* Q_T)N_\xi \right\} \begin{pmatrix} G(\cdot - \xi) \\ G(T - \xi) \end{pmatrix}. \end{aligned} \tag{5.19}$$

By (3.7), (4.2), and (4.3), we obtain

$$\begin{aligned} N_\eta - N_\xi &= N_\eta(I - N_\xi) - (I - N_\eta)N_\xi \\ &= N_\eta \begin{pmatrix} \Gamma \\ \Gamma_T \end{pmatrix} (I - P_\xi)R^{-1}(\Gamma^* Q, \Gamma_T^* Q_T)N_\xi - N_\eta \begin{pmatrix} \Gamma \\ \Gamma_T \end{pmatrix} (I - P_\eta)R^{-1}(\Gamma^* Q, \Gamma_T^* Q_T)N_\xi \\ &= N_\eta \begin{pmatrix} \Gamma \\ \Gamma_T \end{pmatrix} (P_\eta - P_\xi)R^{-1}(\Gamma^* Q, \Gamma_T^* Q_T)N_\xi \\ &= N_\eta \begin{pmatrix} G(\cdot - \eta) \\ G(T - \eta) \end{pmatrix} \Gamma(\eta)(I - P_\xi)R^{-1}(\Gamma^* Q, \Gamma_T^* Q_T)N_\xi. \end{aligned} \tag{5.20}$$

Substitute (5.20) into (5.19); then

$$\begin{aligned} W(t,\eta)W(\eta,\xi) &= G(t-\xi)-G(t-\eta)\Gamma(\eta)(I-P_\xi)R^{-1}(\Gamma^*Q,\Gamma_T^*Q_T)N_\xi\begin{pmatrix}G(\cdot-\xi)\\ G(T-\xi)\end{pmatrix}\\ &\quad -\Gamma(t)(I-P_\eta)R^{-1}(\Gamma^*Q,\Gamma_T^*Q_T)N_\xi\begin{pmatrix}G(\cdot-\xi)\\ G(T-\xi)\end{pmatrix}\\ &= G(t-\xi)-\left\{\int_\xi^\eta+\int_\eta^t\right\}G(t-\sigma)BR^{-1}(\Gamma^*(\sigma)Q,\Gamma_T^*(\sigma)Q_T)N_\xi\\ &\quad\cdot\begin{pmatrix}G(\cdot-\xi)\\ G(T-\xi)\end{pmatrix}d\sigma\\ &= G(t-\xi)-\Gamma(t)(I-P_\xi)R^{-1}(\Gamma^*Q,\Gamma_T^*Q_T)N_\xi\begin{pmatrix}G(\cdot-\xi)\\ G(T-\xi)\end{pmatrix}\\ &= G(t-\xi)-\Gamma(t)(\Phi_\xi^+)^{-1}(I-P_\xi)(\Gamma^*Q,\Gamma_T^*Q_T)\begin{pmatrix}G(\cdot-\xi)\\ G(T-\xi)\end{pmatrix}=W(t,\xi). \end{aligned}$$

(iv) By the definition (5.17), it is clear that $W(t,\xi)$ is strongly continuous in $t\in[\xi,T]$. Combination of this continuity with its uniform boundedness (ii) and its evolutionary property (iii) implies that $W(t,\xi)$ is also strongly continuous in $\xi\in(0,t]$. □

LEMMA 11. *Let $H(t,t)$ be given by* (5.13); *then*

$$H(t,t)=R^{-1}B^*P(t),\qquad t\in[0,T], \tag{5.21}$$

where $P(t)$: $[0,T]\to\mathscr{L}(X)$ is given by

$$\begin{aligned} P(t) &= G^*(T-t)Q_TG(T-t)+\int_t^T G^*(s-t)QG(s-t)\,ds\\ &\quad -(\tilde\Gamma^*(t)Q,\tilde\Gamma_T^*(t)Q_T)N_t\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(I-P_t)R^{-1}(\Gamma^*Q,\Gamma_T^*Q_T)\begin{pmatrix}G(\cdot-t)\\ G(T-t)\end{pmatrix}, \end{aligned} \tag{5.22}$$

where $\tilde\Gamma^*(t)$ and $\tilde\Gamma_T^*(t)$ are defined by (*cf.* (3.12))

$$\begin{aligned} &\tilde\Gamma^*(t)y=\int_t^T G^*(\sigma-t)y(\sigma)\,d\sigma\quad\forall y\in\mathscr{X},\\ &\tilde\Gamma^*(t)y_1=G^*(T-t)y_1\quad\forall y_1\in X. \end{aligned} \tag{5.23}$$

Proof. From (5.13), (5.15), and (4.1), it follows that

$$\begin{aligned} H(t,t) &= \pi(T,t)Q_TG(T-t)+\int_t^T\pi(s,t)QG(s-t)\,ds\\ &= R^{-1}B^*\left[G^*(T-t)Q_TG(T-t)+\int_t^T G^*(s-t)QG(s-t)\,ds\right]\\ &\quad -R^{-1}B^*\int_t^T(\tilde\Gamma^*(t)Q,\tilde\Gamma_T^*(t)Q_T)N_t\begin{pmatrix}G(\cdot-\sigma)\\ G(T-\sigma)\end{pmatrix}\\ &\quad\cdot BR^{-1}B^*G^*(T-\sigma)\,d\sigma\,Q_TG(T-t)\\ &\quad -R^{-1}B^*\int_t^T\int_t^s(\tilde\Gamma^*(t)Q,\tilde\Gamma_T^*(t)Q_T)N_t\begin{pmatrix}G(\cdot-\sigma)\\ G(T-\sigma)\end{pmatrix}\\ &\quad\cdot BR^{-1}B^*G^*(s-\sigma)\,d\sigma\,QG(s-t)\,ds \end{aligned}$$

$$
\begin{aligned}
&= R^{-1}B^*\Big\{G^*(T-t)Q_TG(T-t)+\int_t^T G^*(s-t)QG(s-t)\,ds \\
&\qquad -(\tilde{\Gamma}^*(t)Q,\tilde{\Gamma}_T^*(t)Q_T)N_t\int_t^T\begin{pmatrix}G(\cdot-\sigma)\\ G(T-\sigma)\end{pmatrix} \\
&\qquad\quad \cdot BR^{-1}\Big[B^*G^*(T-\sigma)Q_TG(T-t) \\
&\qquad\quad +\int_\sigma^T B^*G^*(s-\sigma)QG(s-t)\,ds\Big]\,d\sigma\Big\} \\
&= R^{-1}B^*\Big\{G^*(T-t)Q_TG(T-t)+\int_t^T G^*(s-t)QG(s-t)\,ds \\
&\qquad -(\tilde{\Gamma}^*(t)Q,\tilde{\Gamma}_T^*(t)Q_T)N_t\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix} \\
&\qquad\quad \cdot (I-P_t)R^{-1}(\Gamma^*Q,\Gamma_T^*Q_T)\begin{pmatrix}G(\cdot-t)\\ G(T-t)\end{pmatrix}\Big\}.
\end{aligned}
\tag{5.24}
$$

Formula (5.24) indicates that (5.21) and (5.22) are valid. □

LEMMA 12. *Let $P(t)$ and $W(t,\xi)$ be given by* (5.22) *and* (5.17), *respectively. Then the following relation holds*:

$$
P(t)=G^*(T-t)Q_TW(T,t)+\int_t^T G^*(\sigma-t)QW(\sigma,t)\,d\sigma,\qquad t\in(0,T].
\tag{5.25}
$$

Proof. From (5.17) and (4.3), it follows that

$$
\begin{aligned}
&G^*(T-t)Q_TW(T,t)+\int_t^T G^*(\sigma-t)QW(\sigma,t)\,d\sigma \\
&\quad = G^*(T-t)Q_TG(T-t)+\int_t^T G^*(\sigma-t)QG(\sigma-t)\,d\sigma \\
&\qquad -G^*(T-t)Q_T\Gamma(T)(\Phi_t^+)^{-1}(I-P_t)(\Gamma^*Q,\Gamma_T^*Q_T)\begin{pmatrix}G(\cdot-t)\\ G(T-t)\end{pmatrix} \\
&\qquad -\int_t^T G^*(\sigma-t)Q\Gamma(\sigma)(\Phi_t^+)^{-1}(I-P_t)(\Gamma^*Q,\Gamma_T^*Q_T)\begin{pmatrix}G(\cdot-t)\\ G(T-t)\end{pmatrix}d\sigma \\
&\quad = G^*(T-t)Q_TG(T-t)+\int_t^T G^*(\sigma-t)QG(\sigma-t)\,d\sigma \\
&\qquad -(\tilde{\Gamma}^*(t)Q,\tilde{\Gamma}_T^*(t)Q_T)\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(\Phi_t^+)^{-1}(I-P_t)(\Gamma^*Q,\Gamma_T^*Q_T)\begin{pmatrix}G(\cdot-t)\\ G(T-t)\end{pmatrix} \\
&\quad = G^*(T-t)Q_TG(T-t)+\int_t^T G^*(\sigma-t)QG(\sigma-t)\,d\sigma \\
&\qquad -(\tilde{\Gamma}^*(t)Q,\tilde{\Gamma}_T^*(t)Q_T)N_t\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(I-P_t)R^{-1}(\Gamma^*Q,\Gamma_T^*Q_T)\begin{pmatrix}G(\cdot-t)\\ G(T-t)\end{pmatrix}=P(t).
\end{aligned}
$$

□

LEMMA 13. *The following relation holds*:

$$
W(t,\xi)=G(t-\xi)-\int_\xi^t W(t,\sigma)BR^{-1}B^*P(\sigma)G(\sigma-\xi)\,d\sigma,\qquad 0<\xi\leqq t\leqq T,
\tag{5.26}
$$

where $W(t,\xi)$ and $P(t)$ are given by (5.17) *and* (5.22), *respectively.*

Proof. Let the function on the right side of (5.26) be denoted by $\lambda(t, \xi)$. It can be verified that for any $y \in X$, both $g_1(t) = W(t, \xi)y$ and $g_2(t) = \lambda(t, \xi)y$, $t \in [\xi, T]$, are strongly continuous solutions of the following Volterra integral equation:

$$g(t) = G(t-\xi)y - \int_{\xi}^{t} G(t-\sigma)BR^{-1}B^*P(\sigma)g(\sigma)\,d\sigma, \qquad t \in [\xi, T]. \tag{5.27}$$

By the uniqueness of the solution, (5.26) is true. □

LEMMA 14. *Let $P(t)$ be defined by* (5.22). *Then, $P(t)$ is the unique strongly continuous and self-adjoint solution of the following integral Riccati equation*:

$$\begin{aligned} P(t) = {} & G^*(T-t)Q_TG(T-t) + \int_t^T G^*(\sigma-t) \\ & \cdot [Q - P(\sigma)BR^{-1}B^*P(\sigma)]G(\sigma-t)\,d\sigma, \qquad t \in [0, T], \end{aligned} \tag{5.28}$$

where $G(t) = e^{At}$ and $G^*(t) = e^{A^*t}$.

Proof. By Lemma 12, $P(t)$ can be expressed by (5.25). Substitute (5.26) into (5.25); then it follows that

$$\begin{aligned} P(t) &= G^*(T-t)Q_TW(T, t) + \int_t^T G^*(\sigma-t)QW(\sigma, t)\,d\sigma \\ &= G^*(T-t)Q_TG(T-t) + \int_t^T G^*(\sigma-t)QG(\sigma-t)\,d\sigma \\ &\quad - G^*(T-t)Q_T\int_t^T W(T, \sigma)BR^{-1}B^*P(\sigma)G(\sigma-t)\,d\sigma \\ &\quad - \int_t^T G^*(\eta-t)Q\int_t^{\eta} W(\eta, \sigma)BR^{-1}B^*P(\sigma)G(\sigma-t)\,d\sigma\,d\eta \\ &= G^*(T-t)Q_TG(T-t) + \int_t^T G^*(\sigma-t)QG(\sigma-t)\,d\sigma \\ &\quad - \int_t^T G^*(\sigma-t)P(\sigma)BR^{-1}B^*P(\sigma)G(\sigma-t)\,d\sigma \\ &= G^*(T-t)Q_TG(T-t) + \int_t^T G^*(\sigma-t) \\ &\quad \cdot [Q - P(\sigma)BR^{-1}B^*P(\sigma)]G(\sigma-t)\,d\sigma, \qquad t \in [0, T]. \end{aligned}$$

Therefore, $P(t)$ is actually a solution of the integral Riccati equation (5.28). By (5.25) and Lemma 10(iv), $P(t)$ is strongly continuous. In [5], it is proved that the strongly continuous solution of (5.28) is unique. Besides, by transposition it can be seen that $P^*(t)$ is also a solution of (5.28), then the uniqueness implies that $P(t) = P^*(t)$, $t \in [0, T]$. □

It is easy to show that the integral Riccati equation (5.28) is equivalent to the following differential Riccati equation:

$$\begin{aligned} \frac{d}{dt}\langle P(t)x, y\rangle = {} & -\langle P(t)x, Ay\rangle - \langle Ax, P(t)y\rangle - \langle Qx, y\rangle + \langle P(t)BR^{-1}B^*P(t)x, y\rangle \\ & \forall x, y \in \mathscr{D}(A), \quad t \in [0, T], \end{aligned} \tag{5.29}$$

with $P(T) = Q_T$.

Thus, we obtain the following result that shows the synthesis via the Fredholm integral equation and the synthesis via the Riccati equation are equivalent in the standard case.

COROLLARY 5. *If* $A_i = 0$ $(i = 1, \cdots, n)$ *and* $B_j = 0$ $(j = 1, \cdots, m)$, *then the optimal control is given by*

$$u(t) = -R^{-1}B^*P(t)x(t), \qquad t \in [0, T], \tag{5.30}$$

where $P(t)$ *is the unique strongly continuous and self-adjoint solution of the differential Riccati equation* (5.29), *and* $x(\cdot)$ *is the corresponding state trajectory.*

If $Q = 0$ in the quadratic cost functional, then the corresponding linear Fredholm integral equation (4.4) can be solved explicitly to provide an explicit formula for the feedback operator $H(t, \eta)$ in (5.4). Note that in this case,

$$K(t, \sigma) = F^*(T - t)Q_TF(T - \sigma). \tag{5.31}$$

Hence the Fredholm equation (4.4) reduces to

$$M_\xi(t, \sigma) + \int_\xi^T R^{-1}F^*(T - t)Q_TF(T - s)M_\xi(s, \sigma)\, ds = -R^{-1}F^*(T - t)Q_TF(T - \sigma). \tag{5.32}$$

It is easy to show that the solution of (5.32) can be given by an explicit form.

LEMMA 15. *The unique solution of the equation* (5.32) *is given by*

$$M_\xi(t, \sigma) = -R^{-1}F^*(T - t)\sqrt{Q_T}[I + \Lambda(\xi)]^{-1}\sqrt{Q_T}\,F(T - \sigma), \tag{5.33}$$

where

$$\Lambda(\xi) = \int_\xi^T \sqrt{Q_T}\,F(T - t)R^{-1}F^*(T - t)\sqrt{Q_T}\, dt. \tag{5.34}$$

LEMMA 16. *Assume that* $Q = 0$; *then the optimal feedback operator is given by*

$$H(t, n) = R^{-1}F^*(T - t)\sqrt{Q_T}(I + \Lambda(t))^{-1}\sqrt{Q_T}\,G(T - \eta), \tag{5.35}$$

where $\Lambda(t)$ *is given by* (5.34).

Proof. Substitute (5.33) into (4.16) then, in turn, into (5.4); it follows that (5.35) holds by an easy rearrangement. □

COROLLARY 6. *Assume that* $Q = 0$ *so that there is no state integral term in the criterion* (1.5). *Then,* $u(\cdot)$ *is the optimal control if and only if*

$$\begin{aligned} u(t) = -R^{-1}\Bigg\{ & F^*(T - t)\sqrt{Q_T}(I + \Lambda(t))^{-1}\sqrt{Q_T}\,G(T - t)x(t) \\ & + \sum_{i=1}^{n} \int_t^{\min(T, t+h_i)} F^*(T - t)\sqrt{Q_T}(I + \Lambda(t))^{-1}\sqrt{Q_T}\,G(T - \eta)A_ix(\eta - h_i)\, d\eta \\ & + \sum_{i=1}^{n} \int_{\min(t, h_i)}^{h_i} F^*(T - t)\sqrt{Q_T}(I + \Lambda(t))^{-1}\sqrt{Q_T}\,G(T - \eta)A_ix(\eta - h_i)\, d\eta \\ & + \sum_{j=1}^{m} \int_{\min(t, r_j)}^{r_j} F^*(T - t)\sqrt{Q_T}(I + \Lambda(t))^{-1}\sqrt{Q_T}\,G(T - \eta)B_ju(\eta - h_i)\, d\eta \Bigg\}, \\ & \qquad t \in [0, T] \end{aligned} \tag{5.36}$$

where $\Lambda(t)$ *is given by* (5.34).

Remark 3. There is an advantage in the above approach and the obtained closed-loop syntheses, i.e., the dimension of the linear Fredholm integral equation (4.4) is

the same as dim U. In many practical control situations, dim U is finite even though dim X is infinite. Therefore the analytic design of the optimal control reduces to solving a finite-dimensional linear integral equation provided dim $U<\infty$, which can be done by standard numerical methods.

Remark 4. The approach taken in this paper can be applied to solve more complicated problems for retarded functional differential systems with distributed delay in control/state/output variables, neutral differential systems and time-variant systems. It can be applied to tackle quadratic differential game problems as well. In all these possible generalizations, the common link is the semicausal dynamical treatment based on the general Volterra systems.

6. An example. As an example, we consider a control system having partial differential equation model with delay:

$$
\begin{aligned}
&u_t(t,x)=u_{xx}(t,x)+u(t-1,x)+b(x)f(t)+b_1(x)f(t-2), \qquad t\geqq 0, \quad x\in[0,1],\\
&u(t,0)=u(t,1)=0, \qquad t\geqq 0,\\
&u(0,x)=u_0(x)\in L^2(0,1)\triangleq X,\\
&u(\theta,x)=\phi(\theta,x)\in L^2(-1,0;X), \qquad -1\leqq\theta<0, \quad x\in[0,1],\\
&f(\xi)=\psi(\xi)\in L^2(-2,0), \qquad -2\leqq\xi\leqq 0,
\end{aligned}
\tag{6.1}
$$

where we assume that $b(x)$ and $b_1(x)$ belong to $L^2(0,1)$. Let $U=\mathbb{R}$ and $f(\cdot)\in L^2(0,T)=L^2(0,T;\mathbb{R})$ where $T>0$ is finite and fixed.

Set a quadratic cost functional as follows:

$$
J(f)=\|u(T,\cdot)\|_X^2+\int_0^T[\|u(t,\cdot)\|_X^2+|f(t)|^2]\,dt,
\tag{6.2}
$$

and the optimization task is to find a control $f*\in L^2(0,T)$ such that

$$
\min_{f\in L^2(0,T)} J(f)=J(f*).
\tag{6.3}
$$

Define the operator $A:\mathscr{D}(A)\to X$ by

$$
\begin{aligned}
&Au=u_{xx}, \qquad u\in\mathscr{D}(A),\\
&\mathscr{D}(A)=\{u\in X:\ u_{xx}\in X \text{ and } u(0)=u(1)=0\}=H^2(0,1)\cap H_0^1(0,1).
\end{aligned}
\tag{6.4}
$$

Then A is the infinitesimal generator of a self-adjoint and compact semigroup e^{At}, which has an expression:

$$
e^{At}u=\sum_{n=1}^{\infty}e^{\lambda_n t}\langle u,\phi_n\rangle_X\phi_n, \qquad t\geqq 0,
\tag{6.5}
$$

where $\{\lambda_n=-n^2\pi^2:\ n=1,2,\cdots\}=\sigma(A)$ and $\{\phi_n(x)=\sqrt{2}\sin(n\pi x);\ n=1,2,\cdots\}$, as the complete eigenvectors of A, forms an orthonormal basis for X.

On the other hand, define

$$
\begin{aligned}
&A_1=I \qquad \text{(identity on } L^2(0,1)=X),\\
&B=b(\cdot)\in\mathscr{L}(\mathbb{R};X) \quad \text{and} \quad B_1=b_1(\cdot)\in\mathscr{L}(\mathbb{R};X).
\end{aligned}
\tag{6.6}
$$

The task can be rewritten in optimal control formulation described in § 1, with the state equation

$$
\frac{du(t)}{dt}=Au(t)+A_1u(t-1)+Bf(t)+B_1f(t-2)
\tag{6.7}
$$

and the cost functional (1.5), where

$$Q_T = I, \quad Q = I, \quad R = 1. \tag{6.8}$$

The following are steps to find the optimal control.

Step 1. Find the fundamental solution $G(t):\mathbb{R}^+ \to \mathscr{L}(X)$ of (1.9). In this case, by the induction and direct calculation, it is not difficult to obtain the following explicit formula for this $G(t)$:

(6.9)

$$G(t) = \begin{cases} e^{At} + e^{A(t-1)}(t-1) + \cdots + e^{A(t-k)}\dfrac{(t-k)^k}{k!}, & t \in [k, k+1], \quad k = 0, 1, \cdots. \\ 0, & t < 0, \end{cases}$$

Note that for $k = 0$, $G(t) = e^{At}$, $t \in [0, 1]$.

Step 2. Write down the expression of the mild solution of (6.7) as the following Volterra integral system:

$$u(t) = p(t) + \int_0^t F(t-\sigma) f(\sigma)\, d\sigma, \qquad t \in [0, T], \tag{6.10}$$

where $p(t)$ and $F(t)$ are given by

$$p(t) = G(t)u_0 + \int_0^t G(t-\sigma)[\phi(\sigma-1) + b_1\psi(\sigma-2)]\, d\sigma$$

where

$$\phi(t) = \begin{cases} \phi(t, \cdot), & -1 \leqq t < 0 \\ 0, & t \geqq 0 \end{cases} \quad \text{and } \psi(t) = 0 \text{ for } t > 0; \quad \text{and}$$

$$F(t) = G(t)B + G(t-2)B_1 = G(t)b(\cdot) + G(t-2)b_1(\cdot), \qquad t \geqq 0. \tag{6.11}$$

The concrete formula for (6.11) can be obtained by substitution of (6.9) and (6.5) into (6.11).

Step 3. Synthesis equation (4.4). Now we can write the kernel operator function $K(t, \sigma)$ in (2.12) as

$$K(t, \sigma) = F^*(T-t)F(T-\sigma) + \int_t^T F^*(s-t)F(s-\sigma)\, ds \quad \text{for } (t, \sigma) \in [0, T]^2, \tag{6.12}$$

where $F(t)$ is given by (6.11), so $K(t, \sigma)$ is known to us. As mentioned in Corollary 1, we have that $F(t) \equiv 0$ whenever $t < 0$. More precisely, since $F(t) \in \mathscr{L}(\mathbb{R}, X)$ for each $t \in \mathbb{R}$, we have

(6.13)

$$K(t, \sigma) = \langle F(T-t), F(t-\sigma)\rangle_X + \int_t^T \langle F(s-t), F(s-\sigma)\rangle_X\, ds, \qquad (t, \sigma) \in [0, T]^2,$$

which is actually a bivariate scalar continuous function.

Since $R = 1$, the synthesis equation (4.4) now becomes

$$M_\xi(t, \sigma) + \int_\xi^T K(t, s) M_\xi(s, \sigma)\, ds = -K(t, \sigma), \qquad t \in [\xi, T], \quad (\xi, \sigma) \in [0, T]^2, \tag{6.14}$$

where $K(t, \sigma)$ is shown by (6.13). Note that (6.14) is a Fredholm integral equation with scalar unknown function $M_\xi(\cdot, \sigma)$ on $[\xi, T]$ where $(\xi, \sigma) \in [0, T]^2$ are two parameters.

The feature of (6.14) is that this is a scalar integral equation. In general, as long as dim $U < \infty$, the synthesis equation is a finite-dimensional integral equation. Thus

we do not need to solve an infinite-dimensional operator equation such as the Riccati equation.

If in the special case $Q=0$, then the solution of (6.14) is given by the following explicit formula:

$$M_\xi(t,\sigma)=-\left\langle F(T-t),\left[I+\int_\xi^T F(T-t)F^*(T-t)\,dt\right]^{-1}F(T-\sigma)\right\rangle. \tag{6.15}$$

If $Q=I$, then (6.14) can be further solved approximately by numerical methods.

Step 4. Feedback operator $H(t,\sigma)$ and the closed-loop optimal control. From (5.4) and (4.16), we have the following expression. For $(t,\sigma)\in[0,T]^2$,

$$\begin{aligned}
H(t,\sigma)&=\Pi(T,t)G(T-\sigma)+\int_t^T \Pi(s,t)G(s-\sigma)\,ds\\
&=\left[F^*(T-t)+\int_t^T M_t(t,\eta)F^*(T-\eta)\,d\eta\right]G(T-\sigma)\\
&\quad+\int_t^T\left[F^*(s-t)+\int_t^s M_t(t,\eta)F^*(s-\eta)\,d\eta\right]G(s-\sigma)\,ds\\
&=\langle F(T-t),G(T-\sigma)\cdot\rangle+\int_t^T M_t(t,\eta)\langle F(T-\eta),G(T-\sigma)\cdot\rangle\,d\eta\\
&\quad+\int_t^T\langle F(s-t),G(s-\sigma)\cdot\rangle\,ds\\
&\quad+\int_t^T M_t(t,\eta)\int_\eta^T\langle F(s-\eta),G(s-\sigma)\cdot\rangle\,ds\,d\eta.
\end{aligned} \tag{6.16}$$

Finally, substitute (6.16) into (5.3) in Theorem 7, to obtain the closed-loop optimal control as follows:

$$\begin{aligned}
f(t)=-\Bigg[&H(t,t)u(t)+\int_t^{\min(T,t+1)}H(t,\xi)u(\xi-1)\,d\xi\\
&+\int_{\min(t,1)}^1 H(t,\xi)\phi(\xi-1)\,d\xi+\int_{\min(t,2)}^2 H(t,\xi)B_1f(\xi-2)\,d\xi\Bigg],\quad t\in[0,T],
\end{aligned} \tag{6.17}$$

where $H(t,\sigma)$ is shown by (6.16).

Especially, for $t\in[2,T]$, we have the optimal control

$$\begin{aligned}
f(t)=-\Bigg[&\langle F(T-t),G(T-t)u(t)\rangle+\int_t^T\langle F(s-t),G(s-t)u(t)\rangle\,ds\\
&+\int_t^T M_t(t,\eta)\Big(\langle F(T-\eta),G(T-t)u(t)\rangle\\
&+\int_\eta^T\langle F(s-\eta),G(s-t)u(t)\rangle\,ds\Big)\,d\eta\Bigg]\\
-\int_t^{\min(T,t+1)}&\Bigg[\langle F(T-t),G(T-\sigma)u(\sigma-1)\rangle+\int_t^T\langle F(s-t),G(s-\sigma)u(\sigma-1)\rangle\,ds\\
&+\int_t^T M_t(t,\eta)\Big(\langle F(T-\eta),G(T-\sigma)u(\sigma-1)\rangle\\
&+\int_\eta^T\langle F(s-\eta),G(s-\sigma)u(\sigma-1)\rangle\,ds\Big)\,d\eta\Bigg]\,d\sigma.
\end{aligned}$$

Therefore we see that for this example (and also for more general delays), the only essential thing in getting the optimal control is to solve the synthesis equation (6.14) for $M_t(t, \sigma)$, which can be done by means of standard computational methods. □

REFERENCES

[1] R. L. ALFORD AND E. B. LEE, *Sampled data hereditary systems: linear quadratic theory*, IEEE Trans. Automat. Control, (1986), pp. 60-65.

[2] H. T. BANKS, J. A. BURNS, AND E.M. CLIFF, *A comparison of numerical methods for identification and optimization problems involving control systems with delays*, Tech. Report 79-7, Lefschetz Center for Dynamical Systems, Brown University, Providence, RI, 1979.

[3] H. T. BANKS, I. T. ROSEN, AND K. ITO, *A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, ICASE Report 82-31, ICASE, September 1982.

[4] D. CHYUNG AND E. B. LEE, *Linear optimal systems with time delays*, SIAM J. Control, 3 (1966), pp. 548-575.

[5] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, Berlin, New York, 1978.

[6] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, SIAM J. Control, 10 (1972), pp. 298-327.

[7] M. C. DELFOUR, E. B. LEE, AND A. MANITIUS, *F-reduction of the operator Riccati equations for hereditary differential systems*, Automatica, 14 (1978), pp. 385-395.

[8] M. C. DELFOUR, *The linear-quadratic optimal control problem with delays in state and control variables: a state space approach*, SIAM J. Control Optim., 24 (1986), pp. 835-883.

[9] E. FERNANDEZ-BERDAGUER AND E. B. LEE, *Sampled data systems on $M^2 \times L^2$: the quadratic cost formulation for linear delay systems*, in Proc. 22nd IEEE Conference on Decision and Control, 1984, pp. 401-407.

[10] C. A. HARVEY AND G. STEIN, *Quadratic weights for asymptotic regulator properties*, IEEE Trans. Automat. Control, 23 (1978), pp. 378-387.

[11] J. S. GIBSON, *The Riccati integral equations for optimal control problems in a Hilbert space*, SIAM J. Control Optim., 17 (1979), pp. 378-387.

[12] ———, *Linear-quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95-139.

[13] A. ICHIKAWA, *Quadratic control and filtering with delay in control and observation*, Control Theory Center Report, No. 53, University of Warwick, 1977.

[14] H. KOIVO AND E. B. LEE, *Controller synthesis for linear systems with retarded state and control variables and quadratic cost*, Automatica, 8 (1972), pp. 203-208.

[15] R. H. KWONG, *A stability theory for the linear quadratic Gaussian problem for systems with delays in the state, control and observation*, SIAM J. Control Optim., 18 (1980), pp. 49-55.

[16] E. B. LEE AND Y. C. YOU, *Optimal control of bivariate linear Volterra integral type systems*, in Proc. 26th IEEE Conference on Decision and Control, 1987, pp. 721-726.

[17] ———, *Optimal control of two-dimensional linear systems*, in Proc. 26th IEEE Conference on Decision and Control, 1987, pp. 1572-1576.

[18] ———, *Optimal syntheses for infinite dimensional linear delayed state-output systems: semi-causality approach*, J. Appl. Math Optim., (1986), pp. 113-136.

[19] E. B. LEE, *Generalized quadratic optimal controllers for linear hereditary systems*, IEEE Trans. Automat. Control, 25 (1980), pp. 528-531.

[20] A. MANITIUS, *Optimal control of time-lag systems with quadratic performance indexes*, in Proc. 4th IFAC Congress, Session 13, Warsaw, 1969, pp. 16-18.

[21] A. J. PRITCHARD AND D. SALAMON, *The linear quadratic control problem for infinite dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121-144.

[22] A. SCHUMITZKY, *On the equivalence between matrix Riccati equations and Fredholm resolvents*, J. Comput. System Sci., 2 (1968), pp. 76-87.

[23] Y. L. YAO, *On quadratic cost problems for a class of distributed systems with time delay*, Chinese Ann. Math Ser. A, 6 (1985), pp. 50-58. (In Chinese.)

[24] Y. C. YOU, *Closed-loop optimal solution to quadratic boundary control of parabolic systems*, Acta Math. Sinica, 28 (1985), pp. 809-816. (In Chinese.)

[25] Y. C. YOU, *On generator of solution semigroup for linear retarded evolution equation in M^2 space*, Fudan Journal (Natural Sciences), 21 (1982), pp. 163-173. (In Chinese, with English summary.)

[26] ———, *Optimal control problems with hybrid quadratic criteria*, Chinese Ann. Math. Ser. B, 7 (1986), pp. 452-462.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 294-319, March 1990

003

BOUNDARY CONTROLLABILITY OF MAXWELL'S EQUATIONS IN A SPHERICAL REGION*

KATHERINE A. KIME†

Abstract. This paper examines the question of control of electromagnetic fields in a three-dimensional spherical region by means of control currents on the boundary of that region. The necessary theory of divergence-free solutions of the vector wave equation is developed. By use of eigenfunctions of the vector Laplacian in appropriate divergence-free domains and moment problem techniques, sufficient conditions for controllability are set forth.

Key words. control, Maxwell, electromagnetic, boundary control

AMS(MOS) subject classifications. 93C20, 78A40

1. Introduction.

We consider Maxwell's equations

$$\nabla \times H = \frac{\partial E}{\partial t}, \tag{1.1}$$

$$\nabla \times E = -\frac{\partial H}{\partial t}, \tag{1.2}$$

$$\nabla \cdot H = 0, \tag{1.3}$$

$$\nabla \cdot E = 0 \tag{1.4}$$

in Ω the unit ball in R^3, assuming no internal changes or currents. Here $E(x, y, z, t)$ and $H(x, y, z, t)$ are three-dimensional vectors representing the electric and magnetic fields, respectively, $\nabla \times (\ )$ denotes the curl operator, and $\nabla \cdot (\ )$ denotes divergence.

We are interested in the possibility of controlling the fields E, H inside Ω by means of a current $J(\cdot, t)$ flowing tangentially on $\partial\Omega \equiv \Gamma$, the effect of which is described by the boundary condition

$$\bar{n} \times H = -J \quad \text{on} \quad \Gamma \ (\bar{n} \text{ the unit outward normal vector}). \tag{1.5}$$

Thus we may pose the following problem.

CONTROL PROBLEM. Given $T > 0$ and prescribed initial data, find a control current $J(\cdot, t)$ defined on Γ such that the solutions E, H of (1.1)-(1.5) with this initial data also satisfy the terminal condition

$$E(\cdot, T) = H(\cdot, T) = 0.$$

In [15], the control problem for Ω, a circular or rectangular cylinder, was treated, under assumption of invariability of the fields in the axial direction.

Our plan in this paper is as follows. In § 2, we will convert from Maxwell's system to the system of equations for the vector potential associated with the fields, the latter being the vector wave equation with the additional constraint that the solution be divergence-free for all time. It will then be natural to work in certain divergence-free subspaces of

$$\mathscr{L}^2(\Omega) = \{\varphi = (\varphi_1, \varphi_2, \varphi_3) : \varphi_i \in L^2(\Omega),\ i = 1, 2, 3\}.$$

* Received by the editors April 13, 1987; accepted for publication (in revised form) May 26, 1989.

† Department of Mathematics and Statistics, Case Western Reserve University, Cleveland, Ohio 44106.
This research was supported by Office of Army Research contract DAAG 29-80-C-0041, and in part by National Science Foundation grant MCS-P215064 and Air Force Office of Scientific Research grant 85-0283.

We will introduce necessary definitions and give some known, relevant results concerning these subspaces.

In § 3, we establish existence and uniqueness of a weak solution to the system for the vector potential in the case in which the boundary input is in $L^2[0, T; \mathscr{L}^2(\Gamma)]$, here taken as the admissible control space. This is necessary before taking up the control problem itself, which will be done in § 4. By use of divergence-free eigenfunctions of the vector Laplacian, the "multipole fields" [8], [12], we are able to reduce the control problem to a collection of trigonometric moment problems. In § 5, we establish conditions that allow solution of the moment problems.

2. Conversion to potentials; function spaces.

2.1. Conversion to potentials. Let E, H be smooth solutions of (1.1)-(1.5). Then we may see as in [5], [8], and [12], that there exists a smooth vector W with $H = \nabla \times W$, $E = -\partial W/\partial t$, which satisfies

$$\Box W = 0 \quad \text{in } \Omega, \tag{2.1}$$

$$\nabla \cdot W = 0 \quad \text{in } \Omega, \tag{2.2}$$

$$\bar{n} \times (\nabla \times W) = -J \quad \text{on } \Gamma. \tag{2.3}$$

If $H(\cdot, 0) = H_0$, $E(\cdot, 0) \equiv E_0$, $W(\cdot, 0) \equiv W_0$, $\partial W/\partial t(\cdot, 0) \equiv W_1$ then $\nabla \times W_0 = H_0$, $W_1 = -E_0$.

To proceed without undue complication (due mainly to unwieldly limits of integration), we assume E, H are smooth solutions of (1.1)-(1.4) in $\bar{\Omega}^*$, where Ω^* is a slightly larger ball of radius $\sqrt{2}(1+\varepsilon)+\varepsilon$, $\varepsilon > 0$. We may see by well-known arguments that there exists a smooth vector A such that $H = \nabla \times A$ in $\Omega^* \times [0, T]$. With $f(x)$ any smooth function (with $H = (H_1, H_2, H_3)$), we can define

$$A_1 = f(x),$$

$$A_2 = \int_{-(1+\varepsilon)}^{x} H_3(\xi, y, z, t)\, d\xi + K_1(y, z, t),$$

$$A_3 = \int_{-(1+\varepsilon)}^{x} -H_2(\xi, y, z, t)\, d\xi + K_2(y, z, t)$$

where K_1, K_2 are smooth solutions of

$$\frac{\partial K_1}{\partial y} - \frac{\partial K_1}{\partial z} = H_1(-(1+\varepsilon), y, z, t);$$

we take

$$K_1 = 0, \qquad K_2 = \int_{-(1+\varepsilon)}^{y} H_1(-(1+\varepsilon), \eta, z, t))\, d\eta.$$

We now set $A = (A_1, A_2, A_3)$. By (1.2),

$$\nabla \times E = -\frac{\partial H}{\partial t} = -\frac{\partial}{\partial t}(\nabla \times A) = -\nabla \times \left(\frac{\partial A}{\partial t}\right)$$

or

$$\nabla \times \left(E + \frac{\partial A}{\partial t}\right) = 0, \tag{2.4}$$

which implies the existence of ϕ such that

$$E+\frac{\partial A}{\partial t}=-\nabla\phi.$$

We use $\nabla(\ )$ to denote the gradient. By (1.1),

$$\frac{\partial E}{\partial t}=\nabla\times H=\nabla\times(\nabla\times A)=\text{grad div } A-\Delta A,$$

which implies

$$\text{grad div } A-\Delta A=-\frac{\partial}{\partial t}\left(\nabla\phi+\frac{\partial A}{\partial t}\right)$$

or

$$\frac{\partial^2 A}{\partial t^2}-\Delta A=-\nabla\left(\frac{\partial\phi}{\partial t}+\text{div } A\right). \tag{2.5}$$

By (1.4), (2.4), we have

$$\text{div}\left(\nabla\phi+\frac{\partial A}{\partial t}\right)=0,$$

which implies

$$\Delta\phi=-\frac{\partial}{\partial t}(\text{div } A). \tag{2.6}$$

Observe that if W, ψ are defined by

$$W=A+\nabla\chi,\qquad \psi=\phi-\partial\chi/\partial t \tag{2.7}$$

where χ is any smooth scalar field, then

$$\nabla\times W=(\nabla\times A)+(\nabla\times(\nabla\chi))=H,$$

and

$$-\frac{\partial W}{\partial t}-\nabla\psi=-\frac{\partial A}{\partial t}-\frac{\partial}{\partial t}(\nabla\chi)-\nabla\psi+\nabla\left(\frac{\partial\chi}{\partial t}\right)=E.$$

The fields are thus invariant under such transformations, commonly called gauge transformations. We now choose

$$\chi(x, y, z, t)=\int_0^t \phi(x, y, z, s)\,ds+M(x, y, z)$$

where M is the solution of

$$\Delta M=-\text{div } A(x, y, z, 0)\quad \text{in } \Omega,$$
$$M=0\qquad\qquad\qquad\quad \text{on } \Gamma.$$

If $f\in H^K(\Omega)$, then the solution u of

$$\Delta u=f\quad \text{in } \Omega,$$
$$u=0\quad \text{on } \Gamma$$

belongs to $H_0^1(\Omega) \cap H^{2+K}(\Omega)$. Taking $f = -\text{div } A(x, y, z, 0)$ and applying the Sobolev Lemma, we see that M is smooth. Now

$$\psi = \phi - \frac{\partial}{\partial t} \int_0^t \phi(x, y, z, s)\, ds + \frac{dM}{dt} = 0.$$

We have

$$\nabla \cdot W = \nabla \cdot A + \text{div grad } \chi,$$

and

$$\begin{aligned}
\Delta\chi &= \Delta\left(\int_0^t \phi(x, y, z, s)\, ds + M \right) \\
&= \int_0^t \Delta\phi(x, y, z, s)\, ds + \Delta M \\
&= \int_0^t -\frac{\partial}{\partial s} (\text{div } A(x, y, z, s))\, ds - \text{div } A(x, y, z, 0) \\
&= -\text{div } A(x, y, z, t) + \text{div } A(x, y, z, 0) - \text{div } A(x, y, z, 0) \\
&= -\text{div } A(x, y, z, t),
\end{aligned}$$

the third equality from (2.6). Thus

$$\nabla \cdot W = 0. \tag{2.8}$$

W is said to be in the Colomb gauge. Furthermore, since

$$\begin{aligned}
\Box(\nabla\chi) &= \frac{\partial^2}{\partial t^2} (\nabla\chi) - \Delta(\nabla\chi) \\
&= \nabla\left(\frac{\partial^2 \chi}{\partial t^2} \right) - \nabla(\Delta\chi) \\
&= \nabla\left(\frac{\partial \phi}{\partial t} + \text{div } A \right) \\
&= -\Box A,
\end{aligned}$$

the last equality from (2.5), we have

$$\Box W = \Box A + \Box(\nabla\chi) = 0, \tag{2.9}$$

and thus W satisfies (2.1)-(2.3).

It is a matter of straightforward differentiation to see that if, given W satisfying (2.1)-(2.3), we define $H = \nabla \times W$, $E = -\partial W/\partial t$, then E and H satisfy (1.1)-(1.5). For example,

$$\nabla \times H = \nabla \times (\nabla \times W) = -\Delta W = -\frac{\partial^2 W}{\partial t^2} = \frac{\partial E}{\partial t}.$$

2.2. Function spaces. Considering (2.1)-(2.3), we now wish to work in divergence-free subspaces of $\mathscr{L}^2(\Omega)$, as mentioned above. The inner product on $\mathscr{L}^2(\Omega)$ is given by

$$(u, v) = \int_\Omega u \cdot v\, dV,$$

with norm denoted by $\|u\|$.

Let $\mathscr{H}^\ell(\Omega)$ denote the Hilbert space of vectors that belong to $\mathscr{L}^2(\Omega)$ and have derivatives up to and including the order ℓ belonging to $\mathscr{L}^2(\Omega)$. The inner product is

$$(u, v)_\ell = \sum_{|\alpha| \leqq \ell} (D^\alpha u, D^\alpha v) \quad \text{with norm } \|u\|_\ell.$$

We have [9] the following subspace decomposition of $\mathscr{L}^2(\Omega)$:

$$J(\Omega) = \text{closure in } \mathscr{L}^2(\Omega) \text{ of } \{u: u \in C^\infty(\Omega), \operatorname{div} u = 0\},$$
$$\hat{J}(\Omega) = \text{closure in } \mathscr{L}^2(\Omega) \text{ of } \{u: u \in C_0^\infty(\Omega), \operatorname{div} u = 0\},$$
$$G(\Omega) = \{u: u = \nabla \phi, \phi \in H^1(\Omega)\},$$
$$\hat{G}(\Omega) = \{u: u = \nabla \phi, \phi \in H_0^1(\Omega)\},$$
$$J^\ell(\Omega) = \{J(\Omega) \cap \mathscr{H}^\ell(\Omega)\}.$$

Letting the subscript τ denote tangential component, and the subscript n denote normal component, we define

$$J_n^\ell(\Omega) = \{u \in J^\ell(\Omega): u_n|_\Gamma = 0\},$$
$$J_\tau^\ell(\Omega) = \{u \in J^\ell(\Omega): u_\tau|_\Gamma = 0\},$$
$$J_{n,\tau}^\ell(\Omega) = J_n^\ell(\Omega) \cap J_\tau^\ell(\Omega).$$

By $J_n^0(\Omega)$ and by $J_{n,\tau}^0(\Omega)$, we mean $\hat{J}(\Omega)$, and by $J_\tau^0(\Omega)$ we mean $J(\Omega)$. Note that $J_{n,\tau}^\ell(\Omega)$ is dense in $\hat{J}(\Omega)$, $l \geqq 1$.

The following known theorems, Theorems 2.1–2.3, may be found in [9] or [16], and the references therein (see especially [2] and [3]).

Let $E(\Omega) = \{u \in \mathscr{L}^2(\Omega): \operatorname{div} u \in L^2(\Omega)\}$, where $\operatorname{div} u$ is taken in the distributional sense for $u \in \mathscr{L}^2(\Omega)$. E is a Hilbert space when given the scalar product

$$((u, v))_{E(\Omega)} = (u, v)_{\mathscr{L}^2(\Omega)} + (\operatorname{div} u, \operatorname{div} v)_{L^2(\Omega)}.$$

THEOREM 2.1 [16]. *Let Ω be a bounded open set in R^n, with Lipschitz boundary Γ. Then*

$$(\hat{J}(\Omega))^\perp = G(\Omega),$$
$$\hat{J}(\Omega) = \{u \in L^2(\Omega): \operatorname{div} u = 0, \gamma_\nu u = 0\}$$

where $\operatorname{div} u$ *is taken in the distributional sense. Here $\gamma_\nu: E(\Omega) \to H^{-1/2}(\Gamma)$ is a linear continuous operator (the existence of which may be shown) such that*

$$\gamma_\nu u = \text{the restriction of } u \cdot n \text{ to } \Gamma$$

for every $u \in C_0^\infty(\bar{\Omega})$.

THEOREM 2.2 [16], [17]. *Let Ω be a bounded open set of class C^2. Then $\mathscr{L}^2(\Omega) = \hat{J}(\Omega) \oplus K(\Omega) \oplus \hat{G}(\Omega)$, where*

$$K(\Omega) = \{u \in \mathscr{L}^2(\Omega): u = \nabla \phi, \phi \in H^1(\Omega), \Delta \phi = 0\}. \tag{2.10}$$

THEOREM 2.3 [2], [3], [9]. *If Ω is a bounded open set in R^3, with boundary Γ belonging to $C^{\ell+1}$, $\ell \geqq 1$, then the operator curl establishes a one-to-one correspondence between the spaces $J_\tau^\ell(\Omega)$ and $J_n^{\ell-1}(\Omega)$, and also between $J_n^\ell(\Omega)$ and $J^{\ell-1}(\Omega)$, where for all $u \in J_n^\ell(\Omega)$ and $u \in J_\tau^\ell(\Omega)$, we have the estimates*

$$C_1 \|\nabla \times u\|_{\ell-1} \leqq \|u\|_\ell \leqq C_2 \|\nabla \times u\|_{\ell-1}, \tag{2.11}$$

the constants C_1 and C_2 being independent of u.

We will also have need of two additional facts given in [9] concerning these spaces, which we state here as lemmas, giving proofs as the arguments are short.

LEMMA 2.4 [9]. *Let* $J^*(\Omega) = \{u: u \in J_n^2(\Omega), (\nabla \times u)_\tau|_\Gamma = 0\}$. *Then* $J_n^1(\Omega)$ *is the closure of* $J^*(\Omega)$ *in* $\mathscr{H}^1(\Omega)$.

Proof. The indicated closure is clearly contained in $J_n^1(\Omega)$. It also contains $J_n^1(\Omega)$: if $h \in J_n^1(\Omega)$, there exists $\xi \in J(\Omega)$ such that $h = \operatorname{curl} \xi$. As proved in [2], every vector $\xi \in J(\Omega)$ can be approximated by vectors $\xi^m \in J_\tau^1(\Omega)$ in $\mathscr{L}^2(\Omega)$. Let h^m be the solution of the problem $\operatorname{curl} h^m = \xi^m$. Then

$$\|h^m - h\|_1 \leqq C_2 \|\xi^m - \xi\| \to 0.$$

LEMMA 2.5 [9]. *In the Hilbert space* $\hat{J}(\Omega)$, *the operator* $T = \operatorname{curl}\operatorname{curl}$ *with domain* $D(T) = J^*(\Omega)$ *is self-adjoint and positive definite.*

Proof. As $J_{n,\tau}^2(\Omega) \subset J^*(\Omega)$, $J^*(\Omega)$ is dense in $J(\Omega)$. Let u, v be elements of $J^*(\Omega)$. Then

$$(\nabla \times (\nabla \times u), v) = \int_\Omega (\nabla \times (\nabla \times u)) \cdot v \, dV = \int_\Omega (\nabla \times u) \cdot (\nabla \times v) \, dV$$

$$= \int_\Omega u \cdot (\nabla \times (\nabla \times v)) \, dV = (u, \nabla \times (\nabla \times v)).$$

Thus T is symmetric on $J^*(\Omega)$.

Furthermore, the range of T, $R(T)$, is $\hat{J}(\Omega)$: given v an element of $\hat{J}(\Omega)$, by Theorem 2.3 there exists ψ, an element of $J_\tau^1(\Omega)$, such that $\nabla \times \psi = v$. Applying Theorem 2.3 to ψ, we see that there exists ϕ belonging to $J_n^2(\Omega)$ such that $\nabla \times \phi = \psi$. Thus $\phi \in J^*(\Omega)$, and

$$\nabla \times (\nabla \times \phi) = \nabla \times \psi = v.$$

It is well known [19] that if A is a symmetric operator in a Hilbert space H, and $R(A) = H$, then A is self-adjoint. Thus T is self-adjoint. Again employing Theorem 2.3, we have that T is positive definite:

$$(\nabla \times (\nabla \times u), u) = \int_\Omega (\nabla \times u) \cdot (\nabla \times u) \, dV \geqq \frac{1}{C_2} \|u\|_1.$$

We recall the vector identity

$$\nabla \times (\nabla \times u) = \operatorname{grad} \operatorname{div} u - \Delta u. \tag{2.12}$$

From the fact that $\operatorname{div} u = 0$, for $u \in J^*(\Omega)$, we see that $Tu = -\Delta u$ on $J^*(\Omega)$.

Lemma 2.5, along with the fact that the inclusion $\mathscr{H}^1(\Omega) \to \mathscr{L}^2(\Omega)$ is compact implies that the spectrum of $-\Delta$ is discrete and positive and the eigenfunctions of $-\Delta$ in $J^*(\Omega)$ form bases for $\hat{J}(\Omega)$, $J_n^1(\Omega)$, and $J^*(\Omega)$. From Theorem 2.3, we may show that the eigenfunctions are in $J_n^\ell(\Omega)$, ℓ arbitrarily large, and thus, by the Sobolev Lemma, that the eigenfunctions are smooth.

Remark 2.6. Denote the eigenfunctions of $-\Delta$ in $J^*(\Omega)$ by $\{U_\alpha\}$ (the subscript α will be made precise in § 4). If $u \in C^1(\bar{\Omega})$, $y \in C^2(\bar{\Omega})$, then

$$\operatorname{div}(u \times (\nabla \times v)) = (\nabla \times u) \cdot (\nabla \times v) - u \cdot (\nabla \times (\nabla \times v)).$$

Applying the divergence theorem to the left-hand side, we have the following Green identity for vectors:

$$\int_\Omega (\nabla \times u) \cdot (\nabla \times v) \, dV = \int_\Omega u \cdot (\nabla \times (\nabla \times v)) \, dV - \int_\Gamma u \cdot (\bar{n} \times (\nabla \times v)) \, ds. \tag{2.13}$$

Let S denote the set of all linear combinations of the eigenfunctions U_α; S is dense in $\hat{J}(\Omega)$, $J_n^1(\Omega)$, and $J^*(\Omega)$.

Let $u, v \in S$. Then,

$$\int_\Omega (\nabla \times u) \cdot (\nabla \times v)\, dV = -\int_\Omega u\, \Delta v\, dV. \tag{2.14}$$

For fixed v belonging to S, $L(u) = \int_\Omega (\nabla \times u) \cdot (\nabla \times v) + u\, \Delta v\, dV$ may easily be seen to be a bounded linear functional on $J_n^1(\Omega)$. Since $Lu = 0$ for every u belonging to S, by continuity $Lu = 0$ for every $u \in J_n^1(\Omega)$. Thus (2.14) holds for $v \in S$, $u \in J_n^1(\Omega)$. If $u \in J_n^1(\Omega)$, $L(v) = \int_\Omega (\nabla \times u) \cdot (\nabla \times v) + u\, \Delta v\, dV$ defines a bounded linear functional on $J^*(\Omega)$, and again by continuity (2.14) holds for every u belonging to $J_n^1(\Omega)$, v belonging to $J^*(\Omega)$.

3. Solution with L^2 boundary input. In this section we wish to establish the existence and uniqueness of a solution to (2.1)-(2.3), in the case in which $J \in L^2[0, T; \mathscr{L}^2(\Gamma)]$, $T > 0$. The approach will be based on the method of transposition, as put forth in [10a].

Before proceeding, we show that existence and uniqueness of a solution of the system (2.1), (2.2) with the homogeneous boundary condition

$$n \times (\nabla \times W) = 0 \quad \text{on } \Gamma$$

may be obtained, for appropriate initial data, using semigroup techniques. This is of independent but related interest to our result in the case of inhomogeneous boundary data.

THEOREM 3.1. *Given* $\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \in J^*(\Omega) \times J_n^1(\Omega)$, *there exists a unique solution* W *of*

$$\begin{aligned} &\Box W = 0 \\ &\nabla \cdot W = 0 \end{aligned} \quad \text{in } \Omega,$$

$$n \times (\nabla \times W) = 0 \quad \text{on } \Gamma, \tag{3.1}$$

$$W(\cdot, 0) = w_0, \qquad \frac{\partial W}{\partial t}(\cdot, 0) = w_1$$

with $W \in C[0, \infty; J^*(\Omega)] \cap C^1[0, \infty; J_n^1(\Omega)]$.

Proof. We define

$$\hat{H} = \left\{ \begin{bmatrix} u \\ v \end{bmatrix} : u \in J_n^1(\Omega),\ v \in \hat{J}(\Omega) \right\}.$$

With inner product

$$\left(\begin{bmatrix} u \\ v \end{bmatrix}, \begin{bmatrix} y \\ z \end{bmatrix} \right)_{\hat{H}} = \int_\Omega (\nabla \times u) \cdot (\nabla \times y) + v \cdot z\, dV$$

$\hat{H}$ is a Hilbert space. Now we define the operator

$$F = \begin{bmatrix} 0 & I \\ \Delta & 0 \end{bmatrix},$$

with $D(F) = J^*(\Omega) \times J_n^1(\Omega)$. Then, using Remark 2.6, we can see that F is skew-symmetric on $D(F)$:

$$\begin{aligned} \left(F \begin{bmatrix} u \\ v \end{bmatrix}, \begin{bmatrix} y \\ z \end{bmatrix} \right)_{\hat{H}} &= \int_\Omega (\nabla \times v) \cdot (\nabla \times y) + \Delta u \cdot z\, dV \\ &= -\int_\Omega v \cdot \Delta y + (\nabla \times z) \cdot (\nabla \times u)\, dV \\ &= \left(\begin{bmatrix} u \\ v \end{bmatrix}, -F \begin{bmatrix} y \\ z \end{bmatrix} \right)_{\hat{H}}. \end{aligned}$$

By definition, Theorems 2.1, 2.3, and Lemma 2.4, $D(F) \subset \hat{H}$ and is dense in $\hat{H}$. Furthermore, the range of F, $R(F)$, is equal to $\hat{H}$: Given $v \in \hat{J}(\Omega)$, by Theorem 2.3, there exists $\phi \in J_\tau^1(\Omega)$ such that $\nabla \times \psi = -v$. Applying Theorem 2.3 to ψ, we see that there exists $\phi \in J_n^2(\Omega)$ such that $\nabla \times \phi = \psi$. Thus $\phi \in J^*(\Omega)$, and $-(\nabla \times (\nabla \times \phi)) = \Delta \phi = v$.

As in the proof of Lemma 2.5, if A is a symmetric operator on a Hilbert space H and $R(A) = H$, then A is self-adjoint. A straightforward adaptation of the proof of this theorem shows that if A is skew-symmetric, and $R(A) = H$, then A is skew-adjoint. Thus we have shown that F is skew-adjoint on $D(F)$. From [13], F generates a C_0 group of unitary operators $S(t)$ and we now have the desired solution W to (3.1).

As implied by the fact that F is skew-adjoint, if the energy of W is defined as $\|W\|_{\hat{H}}^2$, then for smooth solutions W of the above system,

$$\begin{aligned}\frac{d}{dt}(W, W)_{\hat{H}} &= \frac{d}{dt}\int_\Omega (\nabla \times W) \cdot (\nabla \times W) + W \cdot W \, dV \\ &= 2\int_\Omega \left(\nabla \times \frac{\partial W}{\partial t}\right) \cdot (\nabla \times W) + \frac{\partial W}{\partial t} \cdot \frac{\partial^2 W}{\partial t^2} \, dV \\ &= 2\int_\Omega \frac{\partial W}{\partial t} \cdot \left(-\Delta W + \frac{\partial^2 W}{\partial t^2}\right) dV + 2\int_\Gamma \frac{\partial W}{\partial t} \cdot (n \times (\nabla \times W)) \, ds = 0,\end{aligned}$$

i.e., we have conservation of energy. If we define, as is customary, the electromagnetic energy to be $\int_\Omega H \cdot H + E \cdot E \, dV$, then we have conservation of the electromagnetic energy, since $(W, W)_{\hat{H}} = \int_\Omega H \cdot H + E \cdot E \, dV$.

We now turn to the inhomogeneous system (2.1)–(2.3), starting by stating the following general theorem.

THEOREM 3.2 [10a]. *Let V, H be Hilbert spaces, with*

$$V \subset H, \ V \text{ dense in } H, \ V \text{ separable.}$$

Identifying H with its dual and denoting V^* as the dual of V, we have

$$V \subset H \subset V^*.$$

Let $a(\phi, \psi)$ be a bilinear form on V. Let a be symmetric, and continuous, and assume that there exists $\lambda \in R$ such that

$$a(\psi, \psi) + \lambda \|\psi\|_H^2 \geqq \alpha \|\psi\|_V^2, \ \alpha > 0 \quad \text{for every } \psi \in V, \ t \in [0, T].$$

Then, given $f \in L^2[0, T; H]$, $\delta \in V$, $\kappa \in H$, there exists a unique solution ϕ satisfying

$$\frac{d}{dt}(\phi'(t), \psi) + a(\phi(t), \psi) = (f(t), \psi) \quad \forall \psi \in V \tag{3.2}$$

with $\phi(\cdot, T) = \delta$, $\partial\phi/\partial t(\cdot, T) = \kappa$. The unmarked bracket $(\cdot, \cdot)$ denotes the scalar product in H and $d/dt(\phi'(t), \psi)$ is taken in the distributional sense. In addition,

$$\phi \in C([0, T]; V), \qquad \frac{\partial \phi}{\partial t} \in C([0, T]; H)$$

and the map $\{f, \delta, \kappa\} \to \{\phi, \partial\phi/\partial t\}$ is a continuous linear map $L^2[0, T; H] \times V \times H \to L^2[0, T; V] \times L^2[0, T; H]$.

We may now establish the following theorem.

THEOREM 3.3. *Let*

$$a(\phi, \psi) = \int_\Omega (\nabla \times \phi) \cdot (\nabla \times \psi) \, dV \quad \forall \phi, \psi \in J_n^1(\Omega).$$

Then, given $f \in L^2[0, T; \hat{J}(\Omega)] \equiv Y$, *there exists a unique* ϕ *belonging to* $C[0, T; J_n^1(\Omega)]$, *with* $\partial\phi/\partial t$ *belonging to* $C[0, T; \hat{J}(\Omega)]$, *satisfying*

$$\frac{d}{dt}(\phi'(t), \psi) + a(\phi(t), \psi) = (f(t), \psi) \quad \forall \psi \in J_n^1(\Omega), \tag{3.3}$$

with

$$\phi(\cdot, T) = \frac{\partial \phi}{\partial t}(\cdot, T) = 0. \tag{3.4}$$

Proof. $J_n^1(\Omega) \subset \hat{J}(\Omega)$ by Theorem 2.1, and $J_n^1(\Omega)$ is dense in $\hat{J}(\Omega)$ by definition. By Theorem 2.3, the form a is continuous on $J_n^1(\Omega)$, and satisfies the coercive estimate on $J_n^1(\Omega)$; a is clearly symmetric.

Therefore, taking $V = J_n^1(\Omega)$, $H = \hat{J}(\Omega)$, Theorem 3.2 applies and Theorem 3.3 follows.

Remark 3.4. Equation (3.3) may be interpreted further: as before, denote the eigenfunctions of $-\Delta$ in $J^*(\Omega)$ by $\{U_\alpha\}$. Let $g(t)$ belong to $C_0^\infty[0, T]$, and denote the collection of all products of such $g(t)$, U_α by P. The set Q of all linear combinations of elements of P is dense in $L^2[0, T; \hat{J}(\Omega)]$.

Let $f \in P$, i.e., $f(\cdot, t) = g(t) U_\alpha$. The solution of

$$\Box \tilde{\phi} = f, \qquad \tilde{\phi}_n|_\Gamma = 0,$$

$$\tilde{\phi}(\cdot, T) = \frac{\partial \tilde{\phi}}{\partial t}(\cdot, T) = 0$$

may be obtained by separation of variables; $\tilde{\phi}$ is of the form $\beta(t) U_\alpha$, $\beta(t) \in C_0^\infty[0, T]$ (see § 4). Now

$$(\Box \tilde{\phi}, \psi) = (f, \psi) \quad \forall \psi \in J_n^1(\Omega),$$

$$\left(\frac{\partial^2 \tilde{\phi}}{\partial t^2}, \psi\right) + (-\Delta \tilde{\phi}, \psi) = (f, \psi).$$

From Remark 2.6,

$$(-\Delta \tilde{\phi}, \psi) = a(\tilde{\phi}, \psi) \quad \forall \psi \in J_n^1(\Omega).$$

Thus $\tilde{\phi} = \phi$, where ϕ satisfies (3.2) for $f = g(t) U_\alpha$. Now let $f \in Y$. If ϕ is the corresponding solution to (3.3), then $\Box \phi = f$ in the sense of distributions, on the function space $\mathcal{D} = \{\psi: \psi \in C_0^\infty(\Omega \times [0, T]), \nabla \cdot \psi = 0\}$. We may see this as follows. Let $\{f_k\} \subset Q$ be an approximating sequence for f; solve $\Box \phi_k = f_k$. Let $\psi \in \mathcal{D}$. Then

$$(\Box \phi_k, \psi)_Y = (f_k, \psi)_Y.$$

Integrating by parts,

$$(\phi_k, \Box \psi)_Y = (f_k, \psi)_Y.$$

The map $L: f \to \phi$ is continuous from $Y \to L^2[0, T; J_n^1(\Omega)]$. Thus

$$(\phi, \Box \psi)_Y = (f, \psi)_Y.$$

We now apply the method of transposition [10a]. Let X be the collection of ϕ obtained as f runs over Y; from Theorem 3.3, $\phi(\cdot, T) = \partial\phi/\partial t(\cdot, T) = 0$.

Give X the norm $\|\phi\|_X = \|f\|_Y$; then X is a Hilbert space. Let $S: \phi \to f$; we know S is an isomorphism from X to Y, and that $S\phi = \Box\phi$ in the sense of distributions.

Since S is bounded with respect to X norm, S^* is an isomorphism from Y^* to X^*, the duals of Y and X. Identify X and Y with their duals.

Let $L(\phi)$ be any bounded linear functional on X. Then by the Riesz Representation Theorem there exists $g \in X$ such that

$$L(\phi) = (g, \phi)_X \quad \forall \phi \in X.$$

From the above arguments, there exists $W \in Y$ such that $g = S^* W$. Therefore,

$$L(\phi) = (g, \phi)_X = (S^* W, \phi)_X = (W, S\phi)_Y = \int_0^T \int_\Omega W \cdot S\phi \, dV \, dt.$$

Given $J \in L^2[0, T; \mathscr{L}^2(\Gamma)]$ and $\left[\begin{smallmatrix} w_0 \\ w_1 \end{smallmatrix}\right] \in \hat{J}(\Omega) \times (J_n^1(\Omega))^*$,

$$L(\phi) = \int_0^T \int_\Gamma J \cdot \phi \, ds \, dt + \int_\Omega w_1 \cdot \phi(\cdot, 0) - w_0 \cdot \frac{\partial \phi}{\partial t}(\cdot, 0) \, dV$$

defines a bounded linear functional on X, as in [10a]. Thus we have the following theorem.

THEOREM 3.5. *For J, $\left[\begin{smallmatrix} w_0 \\ w_1 \end{smallmatrix}\right]$ given as above, there exists a unique $W \in Y$ satisfying*

$$\int_0^T \int_\Omega W \cdot \Box\phi \, dV \, dt = \int_0^T \int_\Gamma J \cdot \phi \, ds \, dt + \int_\Omega w_1 \cdot \phi_0 - w_0 \cdot \phi_1 \, dV \tag{3.5}$$

for every $\phi \in X$, where $\phi_0 = \phi(\cdot, 0)$, $\phi_1 = \partial\phi/\partial t(\cdot, 0)$.

We take W to be a weak solution of (2.1)-(2.3). Any smooth solution W^* of (2.1)-(2.3) with initial data w_0^*, w_1^*, corresponding to a smooth solution of (1.1)-(1.5), as in § 2.1, would satisfy (3.5). We show this as follows. On X, define

$$L(\phi) = \int_0^T \int_\Omega W^* \cdot \Box\phi \, dV \, dt - \left[\int_0^T \int_\Gamma J \cdot \phi \, ds \, dt + \int_\Omega w_1^* \cdot \phi_0 - w_0^* \cdot \phi_1 \, dV \right].$$

Since

$$\left| \int_0^T \int_\Omega W^* \cdot \Box\phi \, dV \, dt \right| \leqq \| W^* \|_Y \| \Box\phi \|_Y = \| W^* \|_Y \| \phi \|_X,$$

$L(\phi)$ is bounded on X. Let R be the collection of all solutions ϕ obtained by taking $f \in Q$ (Q as in Remark 3.4).

Let $\hat{\phi} \in X$, and let $\hat{f}$ correspond to $\hat{\phi}$ via (3.2). As Q is dense in Y, there exists $\{f_k\} \subset Q$ such that $\| f_k - \hat{f} \|_Y \to 0$. Solve $\Box\phi_k = f_k$; as before, ϕ_k belongs to R. Then

$$\| \phi_k - \hat{\phi} \|_X = \| f_k - \hat{f} \|_Y \to 0;$$

$\{\phi_k\}$ is an approximating sequence for $\hat{\phi}$ in X. Thus R is dense in X.

On R, $L(\phi_k) = 0$:

$$\begin{aligned}
\int_0^T \int_\Omega W^* \cdot \Box\phi_k \, dV \, dt &= \int_0^T \int_\Omega \left(W^* \cdot \frac{\partial^2}{\partial t^2}(\phi_k) - W^* \cdot \Delta\phi_k \right) dV \, dt \\
&= \int_\Omega w_1^* \cdot \phi_{k,0} - w_0^* \cdot \phi_{k,1} \, dV + \int_0^T \int_\Omega \phi_k \cdot \frac{\partial^2 W^*}{\partial t^2} \, dV \, dt \\
&\quad - \int_0^T \int_\Omega \phi_k \cdot \Delta W^* \, dV \, dt \\
&\quad - \int_0^T \int_\Gamma \phi_k \cdot (n \times (\nabla \times W^*)) \, ds \, dt
\end{aligned}$$

(the last equality from integrating by parts, and applying (2.13))

$$= \int_0^T \int_\Gamma J \cdot \phi_k \, ds \, dt + \int_\Omega w_1^* \cdot \phi_{k,0} - w_0^* \cdot \phi_{k,1} \, dV.$$

Thus, by continuity of L, $L(\phi)=0$ for every $\phi \in X$, and W^* satisfies (3.5) for every $\phi \in X$.

We are interested in controls that are L^2 in time, as we will be using moment problem techniques (see (4.23) and so on). We add that Theorems 3.2 and 3.3 may be improved for inhomogeneous terms $f \in L^1[0, T; \hat{J}(\Omega)]$, leading to an improvement of Theorem 3.5 with controls $J \in L^1[0, T; L^2(\Gamma)]$. We first have Theorem 3.6.

THEOREM 3.6 [10b]. *Let V, H be two real Hilbert spaces with V dense in H, the embedding of V into H continuous, and*

$$V \subset H \subset V^*.$$

Let $a(\phi, \psi)$ be a symmetric continuous bilinear form on V, such that

$$a(\psi, \psi) \geqq \alpha \|\psi\|_V^2, \qquad \alpha > 0.$$

We do not restrict generality by taking

$$a(\psi, \psi) = \|\psi\|_V^2. \tag{3.6}$$

We denote by $A \in L(V, V^*)$ the operator defined by

$$(A\phi, \psi) = a(\phi, \psi).$$

Then, given

$$f \in L^1[0, T; V^*], \quad \phi(0) = \delta \in H, \quad \phi'(0) = k \in V^*,$$

there exists a unique solution ϕ satisfying

$$\phi'' + A\phi = f, \tag{3.7}$$

$$\phi \in L^\infty[0, T; H], \qquad \phi' \in L^\infty[0, T; V^*],$$

$$\phi(\cdot, 0) = \delta, \qquad \phi'(\cdot, 0) = \kappa.$$

We may show that, if we take

$$f \in L^1[0, T; H], \quad \delta \in V, \quad \kappa \in H,$$

then the solution satisfies

$$\phi \in L^\infty[0, T; V], \qquad \phi' \in L^\infty[0, T; H].$$

Multiply (3.7) by ϕ'. We make a small modification of the proof of Theorem 3.6 (in which we multiply (3.7) by $A^{-1}\phi'$, approximate f by regular functions, and follow similar arguments). We have

$$\frac{1}{2}\frac{d}{dt}(\|\phi'\|_H^2 + (A\phi, \phi)) = (f, \phi').$$

Define

$$h(t) = \|\phi'\|_H^2 + \|\phi\|_V^2.$$

Now (using 3.6)

$$\frac{1}{2}\frac{dh}{dt} = \frac{1}{2}\frac{d}{dt}(\|\phi'\|_H^2 + (A\phi, \phi)) \leqq \|f\|_H \|\phi'\|_H,$$

or

$$\frac{dh}{dt} \leqq \|f\|_H (h+1).$$

Then, by Gronwall's Lemma,

$$\|\phi\|^2_{L^\infty[0,T;V]} + \|\phi'\|^2_{L^\infty[0,T;H]} \leqq C \left(\exp \int_0^T \|f\|_H \, dt \right) \times \left(\|\phi(0)\|^2_V + \|\phi'(0)\|^2_H + \int_0^T \|f\|_H \, dt \right).$$

We may reverse the direction of time in Theorem 3.6 and arrive at Theorem 3.7.

THEOREM 3.7. *Let*

$$a(\phi, \psi) = \int_\Omega (\nabla x \phi) \cdot (\nabla x \psi) \, dV \quad \forall \phi, \psi \in J^1_n(\Omega)$$

with

$$A(\phi, \psi) = a(\phi, \psi).$$

Then, given $f \in L^1[0, T; \hat{J}(\Omega)]$, *there exists a unique* ϕ *satisfying*

$$\phi'' + A\phi = f,$$

$$\phi \in L^\infty[0, T; J^1_n(\Omega)], \qquad \phi' \in L^\infty[0, T; \hat{J}(\Omega)],$$

$$\phi(\cdot, T) = \frac{\partial \phi}{\partial t}(\cdot, T) = 0.$$

By extension by continuity, ϕ, ϕ' *are continuous in time.*

Again, by transposition, since $L^\infty[0, T; \hat{J}(\Omega)]$ is the dual of $L^1[0, T; \hat{J}(\Omega)]$, and since the map

$$\phi \to \int_0^T \int_\Gamma J \cdot \phi \, ds \, dt$$

with $J \in L^1[0, T; L^2(\Gamma)]$ is bounded, we have Theorem 3.8.

THEOREM 3.8. *Given*

$$J \in L^1[0, T; L^2(\Gamma)], \qquad \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \in \hat{J}(\Omega) \times (J^1_n(\Omega))^*$$

there exists a unique $W \in L^\infty[0, T; \hat{J}(\Omega)]$ *such that*

$$\int_0^T \int_\Omega W \Box \phi \, dV \, dt = \int_0^T \int_\Gamma J \cdot \phi \, ds \, dt + \int_\Omega (w_1 \cdot \phi_0 - w_0 \cdot \phi_1) \, dV$$

where $\phi_0 = \phi(\cdot, 0)$, $\phi_1 = \partial \phi / \partial t(\cdot, 0)$.

4. Reduction of the control problem to moment problems. Given $\left[\begin{smallmatrix} w_0 \\ w_1 \end{smallmatrix}\right] \in J(\Omega) \times (J^1_n(\Omega))^*$, we now ask if we can find a control J belonging to $L^2[0, T; \mathscr{L}^2(\Gamma)]$ such that the corresponding W given by (3.5) also satisfies

$$W(\cdot, T) = \frac{\partial W}{\partial t}(\cdot, T) = 0.$$

Since W, $\partial W/\partial t$ are not well defined at T, we reformulate: first, take $T_1 > T$, and extend all $J \in L^2[0, T; \mathscr{L}^2(\Gamma)]$ to be identically zero on $[T, T_1]$. This gives a subset $\mathscr{J}$ of $L^2[0, T_1; \mathscr{L}^2(\Gamma)]$. Now replace T by T_1 in Theorem 3.5, and find $J \in \mathscr{J}$ such that the corresponding W, which is thus well-defined on $\Omega \times [T, T_1]$, satisfies $W = 0$ on $[T, T_1]$.

In this section the question of finding the desired control is reduced to the question of finding solutions to a collection of moment problems. Our approach will be similar to that used for the scalar wave equation with L^2 Neumann boundary input [7].

4.1. Equations for J. Now denote the normalized eigenfunctions of $-\Delta$ in $J^*(\Omega)$ by

$$\{U_{nmlk}\}, \quad n = 0, 1, 2, \cdots, \quad m = 0, 1, \cdots, 2n, \quad l = 1, \cdots, \quad k = 1, 2$$

with eigenvalues $(\omega_{nlk})^2$. The U's are basically products of Bessel functions and vector spherical harmonics; they will be given explicitly in § 4.2.

Let $T_1 > T$ and let $h(t)$ be an element of $C_0^\infty[T, T_1]$. Define $h_{nmlk} = h(t) U_{nmlk}$. The set of all such h_{nmlk} is complete in $L^2[T, T_1; \hat{J}(\Omega)]$.

The solution of

$$\Box \phi = h_{nmlk}, \qquad \phi_n|_\Gamma = 0,$$

$$\phi(\cdot, T_1) = \frac{\partial \phi}{\partial t}(\cdot, T_1) = 0$$

is given by

$$\phi(\cdot, t) = \eta(t) U_{nmlk}$$

where $\eta(t)$ satisfies

$$\eta''(t) + (\omega_{nlk})^2 \eta(t) = h(t), \qquad \eta(T_1) = \eta'(T_1) = 0,$$

with $'$ denoting ordinary differentiation. Integrating gives

$$\eta(t) = \frac{1}{\omega_{nlk}} \int_t^{T_1} \sin(\omega_{nlk}(\sigma - t)) h(\sigma)\, d\sigma,$$

$$\eta'(t) = -\int_t^{T_1} \cos(\omega_{nlk}(\sigma - t)) h(\sigma)\, d\sigma;$$

thus

$$\phi(\cdot, t) = \left(\alpha_1 \cos(\omega_{nlk} t) + \frac{1}{\omega_{nlk}} \alpha_2 \sin(\omega_{nlk} t) \right) U_{nmlk}, \tag{4.1}$$

$$\frac{\partial \phi}{\partial t}(\cdot, t) = \left(\alpha_2 \cos(\omega_{nlk} t) - \alpha_1 \omega_{nlk} \sin(\omega_{nlk} t) \right) U_{nmlk} \tag{4.2}$$

where

$$\alpha_1 = \frac{1}{\omega_{nlk}} \int_T^{T_1} \sin(\omega_{nlk} \sigma) h(\sigma)\, d\sigma, \tag{4.3}$$

$$\alpha_2 = -\int_T^{T_1} \cos(\omega_{nlk} \sigma) h(\sigma)\, d\sigma. \tag{4.4}$$

ϕ is clearly in X, and thus for given control J and initial data $\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$, the corresponding weak solution W must satisfy, for every four-tuple $(nmlk)$:

$$\int_0^{T_1}\int_\Omega W\cdot h_{nmlk}\,dV\,dt=\int_\Omega \alpha_1 U_{nmlk}w_1-\alpha_2 U_{nmlk}w_0\,dV$$

$$+\int_0^{T_1}\int_\Gamma J\cdot\left(\alpha_1\cos(\omega_{nlk}t)+\frac{1}{\omega_{nlk}}\alpha_2\sin(\omega_{nlk}t)\right)U_{nmlk}\,dV\,dt. \tag{4.5}$$

Expand

$$w_0=\sum_{n=0}^{\infty}\sum_{m=0}^{2n}\sum_{l=0}^{\infty}\sum_{k=1}^{2}\xi_{0,nmlk}U_{nmlk}, \tag{4.6}$$

$$w_1=\sum_{n=0}^{\infty}\sum_{m=0}^{2n}\sum_{l=0}^{\infty}\sum_{k=1}^{2}\xi_{1,nmlk}Y_{nmlk} \tag{4.7}$$

(assuming $w_1\in\hat{J}(\Omega)$, which is contained in $(J_n^1(\Omega))^*$). Use of (4.6), (4.7), and the orthogonality of the U's allows simplification of (4.5) to obtain

$$\int_T^{T_1}\int_\Omega W\cdot h_{nmlk}\,dV\,dt=\alpha_1\left\{\int_0^T\int_\Gamma J\cdot[\cos(\omega_{nlk}t)U_{nmlk}]\,ds\,dt+\xi_{1,nmlk}\right\}$$

$$+\alpha_2\left\{\int_0^T\int_\Gamma\frac{J\cdot[\sin(\omega_{nlk}t)U_{nmlk}]}{\omega_{nlk}}\,ds\,dt-\xi_{0,nmlk}\right\}. \tag{4.8}$$

Now, if there exists a J satisfying

$$\int_0^T\int_\Gamma J\cdot[\cos(\omega_{nlk}t)U_{nmlk}]\,ds\,dt=-\xi_{1,nmlk}, \tag{4.9}$$

$$\int_0^T\int_\Gamma\frac{J\cdot[\sin(\omega_{nlk}t)U_{nmlk}]}{\omega_{nlk}}\,ds\,dt=\xi_{0,nmlk} \tag{4.10}$$

for every $(nmlk)$, then the corresponding W will satisfy

$$\int_T^{T_1}\int_\Omega W\cdot h_{nmlk}\,dV\,dt=0$$

for every $(nmlk)$, which implies

$$W=0\quad\text{a.e. on }[T,T_1].$$

4.2. Explicit form of eigenfunctions. First, some necessary terminology will be given. Let (r,θ,ϕ) denote the usual spherical coordinates: r is the radial coordinate, $0\leqq r\leqq 1$; θ is the "longitudinal coordinate," $0\leqq\theta\leqq 2\pi$, and ϕ is the "latitudinal coordinate," $0\leqq\phi\leqq\pi$. Thus

$$\int_\Omega dV=\int_0^{2\pi}\int_0^{\pi}\int_0^1 r^2\sin\phi\,dr\,d\phi\,d\theta.$$

The system of basis vectors corresponding to spherical coordinates will be denoted by $\bar{a}_r$, $\bar{a}_\theta$, $\bar{a}_\phi$. The vector $\bar{a}_r$ is the same as $\bar{n}$, the unit outward normal.

Let $Y_{nm}(\theta,\phi)$ denote the spherical harmonics:

$$Y_{nm}(\theta,\phi)=\begin{Bmatrix}(\cos m\theta)P_n^m(\cos\phi), & m=0,1,\cdots,n\\ (\sin[(m-n)\theta])P_n^{m-n}(\cos\phi), & m=n+1,\cdots,2n\end{Bmatrix},\qquad n=0,1,2,\cdots$$

where the $P_n^m(\cos\phi)$ are the Legendre functions [12].

Let $J_{n+1/2}(z)$ be the cylindrical Bessel function of order $n+(\frac{1}{2})$; $J_{n+(1/2)}(kr)$ satisfies

$$\frac{d^2J}{dr^2}+\frac{1}{r}\frac{dJ}{dr}+\left[k^2-\frac{(n+\frac{1}{2})^2}{r^2}\right]J=0. \tag{4.11}$$

Let $j_n(z)=\sqrt{\pi/2z}\,J_{n+(1/2)}(z)$ be the nth spherical Bessel function,

$$\beta_{nl} \text{ be the } l\text{th root of } j_n(\pi\beta)=0, \quad l=1,2,\cdots, \quad \text{and}$$

$$\gamma_{nl} \text{ the } l\text{th root of } \frac{d}{d\gamma}[\pi\gamma j_n(\pi\gamma)]=0, \qquad l=1,2,\cdots.$$

Now, following (essentially) the notation of [12], define

$$\hat{P}_{nm}=Y_{nm}\bar{a}_r, \quad \hat{C}_{nm}=\operatorname{curl}[Y_{nm}\bar{a}_r], \quad \hat{B}_{nm}=\bar{a}_r\times\hat{C}_{nm}.$$

These are referred to as the vector spherical harmonics [1], [4], [12] and are mutually orthogonal. We have

$$\int_\Gamma \hat{P}_{nm}\cdot\hat{P}_{\mu\nu}\,ds=\int_\Gamma \hat{B}_{nm}\cdot\hat{B}_{\nu\mu}\,ds=\int_\Gamma \hat{C}_{nm}\cdot\hat{C}_{\nu\mu}\,ds$$

$$=\frac{4\pi}{\varepsilon_m(2n+1)}\frac{(n+m)!}{(n-m)!}\delta_{m\mu}\delta_{n\nu}\equiv(\tau_{nm})^2\delta_{m\mu}\delta_{n\nu}.$$

Here $\int_\Gamma ds=\int_0^{2\pi}\int_0^{\pi}\sin\phi\,d\phi\,d\theta$, δ denotes the Kronecker delta, and $\varepsilon_m=1$ when $m=0$, two when $m>0$.

We are now ready to give the explicit eigenfunction expressions. Define

$$\begin{aligned}\hat{M}_{nml}&=\operatorname{curl}[j_n(\pi\gamma_{nl}r)Y_{nm}(\theta,\phi)r\bar{a}_r]\\&=\sqrt{n(n+1)}\,j_n(\pi\gamma_{nl}r)\hat{C}_{nm},\end{aligned} \tag{4.12}$$

$$\begin{aligned}\hat{N}_{nml}&=\operatorname{curl}[\operatorname{curl}[j_n(\pi\beta_{nl}r)Y_{nm}(\theta,\phi)r\bar{a}_r]]\\&=\frac{1}{\pi\beta_{nl}r}\left[n(n+1)j_n(\pi\beta_{nl}r)\hat{P}_{nm}+\sqrt{n(n+1)}\left\{\frac{d}{dr}[rj_n(\pi\beta_{nl}r)]\right\}\hat{B}_{nm}\right].\end{aligned} \tag{4.13}$$

If we define

$$(\lambda_{nl})^2\equiv\int_0^1 n(n+1)[j_n(\pi\gamma_{nl}r)]^2r^2\,dr, \tag{4.14}$$

then

$$\int_\Omega \hat{M}_{nml}\cdot\hat{M}_{nml}\,dV=(\lambda_{nl})^2(\tau_{nm})^2,$$

i.e., $\lambda_{nl}\tau_{nm}$ is the normalization constant for $\hat{M}_{nml}$. Similarly, if we define

$$\int_0^1\frac{[n(n+1)j_n(\pi\beta_{nl}r)]^2+n(n+1)\{d/dr[rj_n(\pi\beta_{nl}r)]\}^2}{(\pi\beta_{nl}r)^2}r^2\,dr\equiv(\rho_{nl})^2, \tag{4.15}$$

then

$$\int_\Omega \hat{N}_{nml}\cdot\hat{N}_{nml}\,dV=(\rho_{nl})^2(\tau_{nm})^2, \tag{4.16}$$

and $\rho_{nl}\tau_{nm}$ is the normalization constant for $\hat{N}_{nml}$. Now let

$$C_{nm} = \frac{\hat{C}_{nm}}{\tau_{nm}}, \quad B_{nm} = \frac{\hat{B}_{nm}}{\tau_{nm}}, \quad P_{nm} = \frac{\hat{P}_{nm}}{\tau_{nm}},$$

$$M_{nml} = \frac{\sqrt{n(n+1)}\, j_n(\pi\gamma_{nl} r)}{\lambda_{nl}} C_{nm},$$

$$N_{nml} = \frac{n(n+1)}{\rho_{nl}(\pi\beta_{nl} r)} \left[j_n(\pi\beta_{nl} r) P_{nm} + \frac{1}{\sqrt{n(n+1)}} \left\{ \frac{d}{dr} [r j_n(\pi\beta_{nl} r)] \right\} B_{nm} \right].$$

The M's and N's are the normalized eigenfunctions; $U_{nml1} = M_{nml}$, $U_{nml2} = N_{nml}$. The M's are sometimes referred to as the transverse magnetic fields, and N's as the transverse electric fields. The collection of M's and N's are referred to as the multipole fields [1], [4], [8]. If we replace $\pi\gamma_{nl}$ and $\pi\beta_{nl}$ by k in the expressions for M_{nml} and N_{nml}, and denote these modified functions by M_{nm} and N_{nm}, then

$$M_{nm} = \left(\frac{1}{k}\right) \operatorname{curl} N_{nm}, \qquad N_{nm} = \left(\frac{1}{k}\right) \operatorname{curl} M_{nm}.$$

We may show

$$\begin{aligned} \nabla \times (\nabla \times M_{nml}) &= -\Delta M_{nml} = (\pi\gamma_{nl})^2 M_{nml}, \\ \nabla \times (\nabla \times N_{nml}) &= -\Delta N_{nml} = (\pi\beta_{nl})^2 N_{nml}. \end{aligned} \tag{4.17}$$

Thus $\omega_{nl1} = \pi\gamma_{nl}$, $\omega_{nl2} = \pi\beta_{nl}$. The M's and N's satisfy the boundary condition of zero normal component and zero tangential curl. They are obtained by constructive techniques [12]. It is apparently assumed that these are all the eigenfunctions of the vector Laplacian that are divergence-free and satisfy

$$u_n|_\Gamma = 0, \qquad (\nabla \times u)_\tau|_\Gamma = 0 \tag{4.18}$$

(boundary conditions of $J^*(\Omega)$). We use a theorem of Lamb [18] to verify this, in Appendix A.

We refer to [18] for further discussion of the multipole fields. Also see [11] for a useful summary.

4.3. Obtaining moment problems. From (4.9), (4.10), the desired J must satisfy

$$\int_0^T \int_\Gamma J \cdot \cos(\pi\gamma_{nl} t) M_{nml}\, ds\, dt = -\xi_{1,nml1}, \tag{4.19}$$

$$\int_0^T \int_\Gamma \frac{J \cdot \sin(\pi\gamma_{nl} t)}{\pi\gamma_{nl}} M_{nml}\, ds\, dt = \xi_{0,nml1}, \tag{4.20}$$

$$\int_0^T \int_\Gamma J \cdot \cos(\pi\beta_{nl} t) N_{nml}\, ds\, dt = -\xi_{1,nml2}, \tag{4.21}$$

$$\int_0^T \int_\Gamma \frac{J \cdot \sin(\pi\beta_{nl} t)}{\pi\beta_{nl}} N_{nml}\, ds\, dt = \xi_{0,nml2}. \tag{4.22}$$

Equations (4.19)-(4.22) constitute a moment problem for J in $L^2[0, T; \mathscr{L}^2(\Gamma)]$. We now derive conditions on the Fourier coefficients of J that must hold if J is to satisfy (4.19)-(4.22). These conditions amount to a collection of moment problems; the Fourier coefficients of J must be solutions to these.

Expand

$$J = \sum_{n=1}^{\infty} \sum_{m=0}^{2n} \sigma_{nm}(t) C_{nm} + \pi_{nm}(t) B_{nm} \tag{4.23}$$

(the C's and B's contain only tangential components). Now fix (nm) in (4.19):

$$\int_0^T \int_\Gamma \left[\sum_{p=1}^{\infty} \sum_{q=0}^{2p} \sigma_{pq} C_{pq} + \pi_{pq} B_{pq} \right] \cdot \cos(\pi\gamma_{nl} t) \frac{j_n(\pi\gamma_{nl} r) C_{nm}}{\lambda_{nl}} \, ds \, dt$$

$$= \frac{j_n(\pi\gamma_{nl})}{\lambda_{nl}} \int_0^T \cos(\pi\gamma_{nl} t) \int_\Gamma \left[\sum_{p=1}^{\infty} \sum_{q=0}^{2n} \sigma_{pq} C_{pq} + \pi_{pq} B_{pq} \right] C_{nm} \, ds \, dt$$

$$= \frac{j_n(\pi\gamma_{nl})}{\lambda_{nl}} \int_0^T \cos(\pi\gamma_{nl} t) \sigma_{nm}(t) \, dt.$$

Thus, for every $l = 0, 1, 2, \cdots$, $\sigma_{nm}(t)$ must satisfy

$$\frac{\sqrt{n(n+1)} j_n(\pi\gamma_{nl})}{\lambda_{nl}} \int_0^T \cos(\pi\gamma_{nl} t) \sigma_{nm}(t) \, dt = -\xi_{1,nml1}. \tag{4.24}$$

Similarly, from (4.20), for every $l = 0, 1, 2, \cdots$, $\sigma_{nm}(t)$ must satisfy

$$\frac{\sqrt{n(n+1)} j_n(\pi\gamma_{nl})}{\lambda_{nl}(\pi\gamma_{nl})} \int_0^T \sin(\pi\gamma_{nl} t) \sigma_{nm}(t) \, dt = \xi_{0,nml1}. \tag{4.25}$$

Now treating the pair (4.21), (4.22) similarly, we have

$$\int_0^T \int_\Gamma J \cdot \frac{\cos(\pi\beta_{nl} t)}{\rho_{nl}(\pi\beta_{nl} r)} \left[j_n(\pi\beta_{nl} r) P_{nm} + \frac{\{d/dr[r j_n(\pi\beta_{nl} r)] B_{nm}\}}{\sqrt{n(n+1)}} \right] ds \, dt$$

$$= \frac{1}{\rho_{nl}} \int_0^T \cos(\pi\beta_{nl} t) \int_\Gamma \left[\sum_{p=1}^{\infty} \sum_{q=1}^{2p} \sigma_{pq} C_{pq} + \pi_{pq} B_{pq} \right] \cdot \frac{j_n'(\pi\beta_{nl}) B_{nm}}{\sqrt{n(n+1)}} \, ds \, dt$$

(since $d/dr[r j_n(\pi\beta_{nl} r)] = j_n(\pi\beta_{nl} r) + \pi\beta_{nl} r j_n'(\pi\beta_{nl} r) = \pi\beta_{nl} j_n'(\pi\beta_{nl})$ at $r = 1$)

$$= \frac{j_n'(\pi\beta_{nl})}{\rho_{nl}\sqrt{n(n+1)}} \int_0^T \cos(\pi\beta_{nl} t) \pi_{nm}(t) \, dt.$$

Thus, for every $l = 0, 1, 2, \cdots$, $\pi_{nm}(t)$ must satisfy

$$\frac{\sqrt{n(n+1)} j_n'(\pi\beta_{nl})}{\rho_{nl}} \int_0^T \cos(\pi\beta_{nl} t) \pi_{nm}(t) \, dt = -\xi_{1,nml2}. \tag{4.26}$$

Similarly, from (4.22), for every $l = 0, 1, 2, \cdots$, $\pi_{nm}(t)$ must satisfy

$$\frac{\sqrt{n(n+1)} j_n'(\pi\beta_{nl})}{\pi\beta_{nl} \cdot \rho_{nl}} \int_0^T \sin(\pi\beta_{nm} t) \pi_{nm}(t) \, dt = \xi_{0,nml2}. \tag{4.27}$$

Equations (4.24) and (4.25) may be put in the equivalent form:

$$\int_0^T \exp[i\pi\gamma_{nl} t] \sigma_{nm}(t) \, dt = \frac{-\xi_{1,nml1} + i\pi\gamma_{nl}\xi_{0,nml1}}{\{\sqrt{n(n+1)} j_n(\pi\gamma_{nl})\}/(\lambda_{nl})} \equiv a_{nml}, \tag{4.28}$$

$$\int_0^T \exp[-i\pi\gamma_{nl} t] \sigma_{nm}(t) \, dt = \frac{-\xi_{1,nml1} - i\pi\gamma_{nl}\xi_{0,nml1}}{\{\sqrt{n(n+1)} j_n(\pi\gamma_{nl})\}/(\lambda_{nl})} \equiv b_{nml}. \tag{4.29}$$

Equations (4.26) and (4.27) may be put in the equivalent form:

$$\int_0^T \exp[i\pi\beta_{nl} t] \pi_{nm}(t) \, dt = \frac{-\xi_{1,nml2} + i\pi\beta_{nl}\xi_{0,nml2}}{\{\sqrt{n(n+1)} j_n'(\pi\beta_{nl})\}/(\rho_{nl})} \equiv c_{nml}, \tag{4.30}$$

$$\int_0^T \exp[-i\pi\beta_{nl} t] \pi_{nm}(t) \, dt = \frac{-\xi_{1,nml2} - i\pi\beta_{nl}\xi_{0,nml2}}{\{\sqrt{n(n+1)} j_n'(\pi\beta_{nl})\}/(\rho_{nl})} \equiv d_{nml}. \tag{4.31}$$

Equations (4.28) and (4.29) constitute a moment problem for $\sigma_{nm}(t)$; (4.30) and (4.31) constitute a moment problem for $\pi_{nm}(t)$ (the reader may refer to, e.g., [7] and [14] for further discussions of moment problems). In § 5, we will show that, under certain conditions, solutions to these moment problems exist.

5. Solutions of moment problems and sufficient conditions on initial data. In § 5.1, we prove that, under summability assumptions on the a's, b's, c's, and d's, the moment problems (4.28), (4.29), and (4.30), (4.31) have solutions. These solutions may be used in the role of the Fourier coefficients of J, and, under additional summability assumptions, the J so defined does belong to $L^2[0, T; \mathscr{L}^2(\Gamma)]$ and satisfies (4.19)-(4.22). In § 5.2, we give sufficient conditions on the initial data w_0, w_1 to ensure satisfaction of all the summability assumptions, culminating in Theorem 5.6.

5.1. Solutions to moment problems.

PROPOSITION 5.1. *For $n = 1, 2, 3, \cdots$, the intervals between the consecutive positive roots of $d/d\gamma[\gamma j_n(\pi\gamma)] = 0$ are bounded below by π, and decrease monotonically to π as $l \to \infty$:*

$$\pi\gamma_{n2} - \pi\gamma_{n1} > \cdots > \pi\gamma_{n,l+1} - \pi\gamma_{n,l} > \cdots \to \pi.$$

Proof. As in § 4.3,

$$\frac{d}{dy}[\gamma j_n(\pi\gamma)] = j_n(\pi\gamma) + \pi\gamma j_n'(\pi\gamma) = 0 \quad \text{for } \gamma = \gamma_{nl}, \quad \text{i.e.,} \tag{5.1}$$

$$j_n(\pi\gamma_{nl}) + \pi\gamma_{nl} j_n'(\pi\gamma_{nl}) = 0.$$

Define

$$f(r) = \frac{d}{dr}[r j_n(\pi\gamma_{nl} r)] = j_n(\pi\gamma_{nl} r) + r\frac{d}{dr}[j_n(\pi\gamma_{nl} r)].$$

From § 4.2,

$$j_n(\pi\gamma_{nl} r) = \frac{1}{\sqrt{2\gamma_{nl} r}} J_{n+(1/2)}(\pi\gamma_{nl} r).$$

Thus,

$$\begin{aligned} f(r) &= \frac{1}{\sqrt{2\gamma_{nl} r}} J_{n+(1/2)}(\pi\gamma_{nl} r) + r \cdot \frac{d}{dr}\left[\frac{1}{\sqrt{2\gamma_{nl} r}} J_{n+(1/2)}(\pi\gamma_{nl} r)\right] \\ &= \frac{1}{\sqrt{2\gamma_{nl}}}\left[\frac{J_{n+(1/2)}(\pi\gamma_{nl} r)}{\sqrt{r}} + r \cdot \left\{-\frac{J_{n+(1/2)}(\pi\gamma_{nl} r)}{2r\sqrt{r}} + \frac{\pi\gamma_{nl}}{\sqrt{r}} J'_{n+(1/2)}(\pi\gamma_{nl} r)\right\}\right] \\ &= \frac{1}{\sqrt{2\gamma_{nl} r}}\left[\frac{1}{2} J_{n+(1/2)}(\pi\gamma_{nl} r) + \pi\gamma_{nl} r J'_{n+(1/2)}(\pi\gamma_{nl} r)\right]. \end{aligned}$$

From (5.1), $f(1) = 0$. Therefore, our equation assumes the form

$$\frac{1}{\sqrt{2\gamma_{nl}}}\left[\frac{1}{2} J_{n+(1/2)}(\pi\gamma_{nl}) + \pi\gamma_{nl} J'_{n+(1/2)}(\pi\gamma_{nl})\right] = 0. \tag{5.2}$$

We now turn to the work of Graham [6], in which he defines, for real x, the positive solutions of

$$\frac{1}{\sqrt{\omega}}\left[\frac{1}{2} J_x(\omega) + \omega J'_x(\omega)\right] = 0 \tag{5.3}$$

as τ_{xl}, $l=1, 2, \cdots$ and the zeros of $J_x(\omega)$ to be g_{xl}, $l=1, 2, \cdots$ (we use g_{xl} for Graham's j_{xl}). Thus from (5.2), and the definition of j_n, we have

$$\pi\gamma_{nl} = \tau_{n+(1/2),l}, \qquad \pi\beta_{nl} = g_{n+(1/2),l}.$$

Reference [6] contains the following result (Lemma 3.1). If $x > \frac{1}{2}$, then the lengths of the intervals between successive elements of the sequence

$$\cdots \tau_{x,l-1} < g_{x,l-1} < \tau_{xl} \cdots$$

decrease monotonically to their asymptotic value of $\pi/2$,

$$g_{x1} - \tau_{x1} > \tau_{x2} - g_{x1} > \cdots \to \pi/2.$$

If $x = \frac{1}{2}$, then all these intervals are exactly $\pi/2$ long.

The proposition follows.

Remarks 5.2. (1) Graham's Lemma 3.1 of [6] and the work of Watson (see [6]) give a lower bound of π on the spacings of the $\pi\beta_{nl}$'s.

(2) Lemma 3.1 of [6] ensures that $d/dr[rj_n(\pi\beta_{nl}r)]$ and $j_n(\pi\gamma_{nl}r)$ do not vanish at $r=1$, a fact that was implicitly assumed true in the derivations of (4.28)-(4.31). We see this as follows. Clearly, $j_n(\pi\gamma_{nl}) \neq 0$. Furthermore, if we let

$$f(r) = \frac{d}{dr}[rj_n(\pi\beta_{nl}r)]$$

then, computing as before, with $x = n + (\frac{1}{2})$,

$$f(r) = \frac{1}{\sqrt{2\beta_{nl}r}}\left[\frac{1}{2}J_x(\pi\beta_{nl}r) + \pi\beta_{nl}rJ'_x(\pi\beta_{nl}r)\right].$$

Since

$$\frac{1}{2}J_x(\omega) + \omega J'_x(\omega) = 0$$

for $\omega = \tau_{xl}$, and τ_{xl} and g_{xl} $(=\pi\beta_{nl})$ are separated by a gap of at least $\pi/2$, we see that $f(1) \neq 0$.

(3) Separation of variables for the scalar Helmholtz equation $\Delta U + \lambda U = 0$, in the unit ball in R^3, with the homogeneous elastic boundary condition $\partial U/\partial n + U = 0$ gives rise to (5.3).

THEOREM 5.3. *Let $n \geqq 1$, $0 \leqq m \leqq 2n$. Assume the sequences $\{a_{nml}\}$, $\{b_{nml}\}$, $\{c_{nml}\}$, and $\{d_{nml}\}$, $l=1, 2, \cdots$, are square summable. Then, given $T > 2$:*

(i) *The moment problem* (4.28), (4.29) *has a solution $\sigma_{nm}(t)$, with*

$$\Lambda_1 \sum_{l=1}^{\infty} (|a_{nml}|^2 + |b_{nml}|^2) \leqq \|\hat{\sigma}_{nm}(t)\|^2_{L^2[0,T]} \leqq \Pi_1 \sum_{l=1}^{\infty} (|a_{nml}|^2 + |b_{nml}|^2) \tag{5.4}$$

where Λ_1, Π_1 are constants independent of nm and the particular sequences a_{nml} and b_{nml}.

(ii) *The moment problem* (4.30), (4.31) *has a solution $\hat{\pi}_{nm}(t)$, with*

$$\Lambda_2 \sum_{l=1}^{\infty} (|c_{nml}|^2 + |d_{nml}|^2) \leqq \|\hat{\pi}_{nm}(t)\|_{L^2[0,T]} \leqq \Pi_2 \sum_{l=1}^{\infty} (|c_{nml}|^2 + |d_{nml}|^2) \tag{5.5}$$

where Λ_2 and Π_2 are constants analogous to Λ_1 and Π_1.

Proof. By Proposition 5.1,

$$\frac{2\pi}{\min_l (\pi\gamma_{n,l+1} - \pi\gamma_{nl})} = 2.$$

Thus, if $T > 2$, a result of Ingham, as in [7], shows that hypotheses of a theorem of Boas may be satisfied. Therefore, the solution $\hat{\sigma}_{nm}(t)$ exists, and (5.4) holds.

Similarly, the lower bound of π on the spacings of the $\pi\beta_{nl}$'s gives the existence of $\hat{\pi}_{nm}(t)$, and (5.5).

THEOREM 5.4. *Assume*

$$\sum_{n=1}^{\infty} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} |a_{nml}|^2 < \infty, \qquad \sum_{n=1}^{\infty} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} |c_{nml}|^2 < \infty.$$

If we define

$$J = \sum_{n=1}^{\infty} \sum_{m=0}^{2n} \hat{\sigma}_{nm}(t) C_{nm} + \hat{\pi}_{nm}(t) B_{nm},$$

using the solutions $\hat{\sigma}_{nm}(t)$, $\hat{\pi}_{nm}(t)$ *from Theorem* 5.3, *then* $J \in L^2[0, T; \mathscr{L}^2(\Gamma)]$, and J *satisfies* (4.19)-(4.22).

Proof. Note first that $|a_{nml}| = |b_{nml}|$, $|c_{nml}| = |d_{nml}|$:

$$\left\| \sum_{n=n_1}^{n=n_2} \sum_{m=0}^{2n} \{\hat{\sigma}_{nm}(t) C_{nm} + \hat{\pi}_{nm}(t) B_{nm}\} \right\|_{L^2[0,T;\mathscr{L}^2(\Gamma)]}$$

$$= \sum_{n=n_1}^{n=n_2} \sum_{m=0}^{2n} \{\|\hat{\sigma}_{nm}(t)\|_{L^2[0,T]} + \|\hat{\pi}_{nm}(t)\|_{L^2[0,T]}\}$$

$$\leqq 2\Pi_1 \sum_{n=n_1}^{n=n_2} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} |a_{nml}|^2 + 2\Pi_2 \sum_{n=n_1}^{n=n_2} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} |c_{nml}|^2$$

$$\to 0 \quad \text{as } n_1, n_2 \to 0.$$

The inequality follows from (5.4), (5.5).

The fact that J satisfies (4.19)-(4.22) follows from the discussion preceding (4.28)-(4.31).

5.2. Sufficient conditions on initial data. We first simplify some notation. Let

$$\mu_{nml} \equiv \xi_{0,nml1}, \qquad \zeta_{nml} \equiv \xi_{1,nml1},$$

$$\nu_{nml} \equiv \xi_{0,nml2}, \qquad \eta_{nml} \equiv \xi_{1,nml2}.$$

Thus

$$w_0 = \sum_{n=0}^{\infty} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} \mu_{nml} M_{nml} + \nu_{nml} N_{nml},$$

$$w_1 = \sum_{n=0}^{\infty} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} \zeta_{nml} M_{nml} + \eta_{nml} N_{nml}.$$

We will have need of simplified expressions for the constants λ_{nl} and ρ_{nl}, defined in (4.14), (4.15). These expressions are:

$$(\lambda_{nl})^2 = \frac{n(n+1)}{2}\left[1 - \frac{n(n+1)}{(\pi\gamma_{nl})^2}\right] j_n^2(\pi\gamma_{nl}), \tag{5.6}$$

$$(\rho_{nl})^2 = \frac{n(n+1)}{2} j_{n+1}^2(\pi\beta_{nl}). \tag{5.7}$$

They are derived in Appendix B.

THEOREM 5.5.

(5.8) (i) $$\sum_{l=1}^{\infty} |a_{nml}|^2 < \sum_{l=1}^{\infty} (|\zeta_{nml}|^2 + |\pi\gamma_{nl}\mu_{nml}|^2),$$

(5.9) (ii) $$\sum_{l=1}^{\infty} |c_{nml}|^2 \leqq \sum_{l=1}^{\infty} (|\eta_{nml}|^2 + |\pi\beta_{nl}\nu_{nml}|^2)$$

(*assume the right-hand sides of* (5.8) *and* (5.9) *are finite*; *Theorem* 5.6 *will address this assumption*).

Proof.

$$|a_{nml}|^2 = \left| \frac{-\zeta_{nml} + i\pi\gamma_{nl}\mu_{nml}}{\{\sqrt{n(n+1)}\, j_n(\pi\gamma_{nl})\}/(\lambda_{nl})} \right|^2$$

$$\leqq 2 \frac{\{|\zeta_{nml}|^2 + |\pi\gamma_{nl}\mu_{nml}|^2\}}{R_{nl}^2}$$

where R_{nl} denotes the denominator of a_{nml}.

$$|R_{nl}| = \frac{\sqrt{n(n+1)}\,|j_n(\pi\gamma_{nl})|}{\left(\frac{n(n+1)}{2}\left(1 - \frac{n(n+1)}{(\pi\gamma_{nl})^2}\right) j_n^2(\pi\gamma_{nl})\right)^{1/2}}.$$

As $l \to \infty$, $1 - (n(n+1)/(\pi\gamma_{nl})^2) \to 1$. Therefore,

$$|R_{nl}| > \sqrt{2}, \text{ and (i) follows.}$$

Proceeding similarly for the proof of (ii):

$$|c_{nml}|^2 = \left| \frac{-\eta_{nml} + i\pi\beta_{nl}\nu_{nml}}{\{\sqrt{n(n+1)}\, j_n'(\pi\beta_{nl})\}/(\rho_{nl})} \right|^2$$

$$\leqq 2 \frac{\{|\eta_{nml}|^2 + |\pi\beta_{nl}\nu_{nml}|^2\}}{Q_{nl}^2}$$

where Q_{nl} denotes the denominator of c_{nml}. From identity (2) (Appendix B), for the spherical Bessel functions, we have

$$(2n+1) j_n'(\pi\beta_{nl}) = n j_{n-1}(\pi\beta_{nl}) - (n+1) j_{n+1}(\pi\beta_{nl}).$$

Employing identity (1) (Appendix B), we have

$$(2n+1) j_n'(\pi\beta_{nl}) = n j_{n-1}(\pi\beta_{nl}) + (n+1) j_{n-1}(\pi\beta_{nl}).$$

Therefore

$$j_n'(\pi\beta_{nl}) = j_{n-1}(\pi\beta_{nl}).$$

We now have

$$|Q_{nl}| = \frac{\sqrt{n(n+1)}\,|j_n'(\pi\beta_{nl})|}{\left[\frac{n(n+1)}{2}\right]^{1/2} |j_{n-1}(\pi\beta_{nl})|} = \sqrt{2},$$

and (ii) follows.

THEOREM 5.6. *Let* $T>2$, *and let* $\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \in J_n^1(\Omega) \times \hat{J}(\Omega)$. *Then*

$$\text{(5.10)} \quad \text{(i)} \qquad \sum_{n=1}^{\infty} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} |\pi\gamma_{nl}\mu_{nml}|^2 < \infty,$$

$$\text{(5.11)} \qquad \sum_{n=1}^{\infty} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} |\pi\beta_{nl}\nu_{nml}|^2 < \infty.$$

$$\text{(ii)} \qquad \sum_{n=1}^{\infty} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} |\zeta_{nml}|^2 < \infty, \qquad \sum_{n=1}^{\infty} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} |\eta_{nml}|^2 < \infty.$$

(iii) *There exists* $J \in L^2[0, T, \mathscr{L}^2(\Gamma)]$, *which satisfies* (4.19)-(4.22), *and hence the corresponding* W, *given by Theorem* 3.5, *satisfies* $W=0$ *almost everywhere on* $[T, T_1]$.

Proof. Part (ii) follows directly from the fact that $w_0 \in J(\hat{\Omega})$. Turning to (i), we have

$$\text{(5.12)} \qquad \begin{aligned} \mu_{nml} &= \int_\Omega w_0 \cdot M_{nml}\, dV = \frac{-1}{(\pi\gamma_{nl})^2} \int_\Omega w_0 \cdot \Delta M_{nml}\, dV \\ &= \frac{1}{(\pi\gamma_{nl})^2} \int_\Omega (\nabla \times w_0) \cdot (\nabla \times M_{nml})\, dV. \end{aligned}$$

As may be deduced from remarks in § 4.2,

$$\nabla \times M_{nml} = (\pi\gamma_{nl}) N^*_{nml}$$

where N^*_{nml} denotes the function that results from replacing $\pi\beta_{nl}$ by $\pi\gamma_{nl}$ in N_{nml}. The collection of N^*_{nml} is an orthonormal set in $J(\Omega)$. The "Fourier coefficient" of $(\nabla \times w_0)$ with respect to N^*_{nml} is

$$\int_\Omega (\nabla \times w_0) \cdot \frac{(\nabla \times M_{nml})}{(\pi\gamma_{nl})}\, dV,$$

which we see from (5.12) is equal to $\pi\gamma_{nl}\mu_{nml}$. Since the "Fourier coefficients" of $(\nabla \times w_0)$ with respect to the N^*_{nml}'s must be square summable, by Bessel's inequality,

$$\sum_{n=1}^{\infty} \sum_{m=0}^{2n} \sum_{l=1}^{\infty} |\pi\gamma_{nl}\mu_{nml}|^2 < \infty.$$

Similar arguments apply to the terms $\pi\beta_{nl}\nu_{nml}$. Thus (i) is proved.

Now (i), (ii) and Theorem 5.5 imply that the hypotheses of Theorem 5.4 are satisfied, and (iii) follows.

We remark that, by time reversibility of the wave equation, we are thus able to steer from any given terminal data $\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \in J_n^1(\Omega) \times J^*(\Omega)$ to any given terminal data $\begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$ in the same space. We also note that the solution W given by Theorem 3.5 belongs to $L^2[0, T; \hat{J}(\Omega)]$ and there is no guarantee that the electromagnetic energy

$$\int_\Omega (\nabla \times W) \cdot (\nabla \times W) + \frac{\partial W}{\partial t} \cdot \frac{\partial W}{\partial t}\, dV = \int_\Omega H \cdot H + E \cdot E\, dV$$

remains finite, even when starting from finite energy states belonging to $J_n^1(\Omega) \times \hat{J}(\Omega)$. An analogous situation occurs in the case of control of the scalar wave equation via L^2 Neumann boundary controls in the unit ball Ω [7].

Appendix A. Here we show that the collection of M_{nml}'s and N_{nml}'s are all the eigenfunctions of the vector Laplacian that are divergence-free and satisfy

$$\text{(A1)} \qquad u_n|_\Gamma = 0, \ (\nabla \times u)_\tau|_\Gamma = 0,$$

i.e., are all the eigenfunctions of $-\Delta$ in $J^*(\Omega)$.

Suppose not, i.e., suppose there exists a smooth P that satisfies

$$-\Delta P = k^2 P, \qquad \operatorname{div} P = 0,$$

and (A1), and that is orthogonal to all M_{nml}, N_{nml}. Let

$$j_n(\pi\gamma_{nl}r)\,Y_{nm} = \psi_{nml}, \qquad j_n(\pi\beta_{nl}r)\,Y_{nm} = \chi_{nml}.$$

Then

$$M_{nml} = \nabla \times (\psi_{nml} r\bar{a}_r), \qquad N_{nml} = \nabla \times (\nabla \times (\chi_{nml} r\bar{a}_r),$$

omitting normalization constants for convenience. By assumption,

$$\begin{aligned} 0 = \int_\Omega P \cdot N_{nml}\, dv &= \int_\Omega P \cdot \nabla \times (\nabla \times (\chi_{nml} r\bar{a}_r))\, dV \\ &= \left\{ \int_\Omega (\nabla \times P) \cdot (\nabla \times (\chi_{nml} r\bar{a}_r)\, dV + \int_\Gamma P \cdot \bar{a}_r \times (\nabla \times \chi_{nml} r\bar{a}_r)\, ds \right\} \\ &= \left\{ \int_\Omega (\chi_{nml} r\bar{a}_r) \cdot (k^2 P)\, dV + \int_\Gamma \chi_{nml} r\bar{a}_r \cdot (\bar{a}_r \times (\nabla \times P))\, ds \right\}. \end{aligned}$$

The last equality follows from the fact that

$$\nabla \times (\chi_{nml} r\bar{a}_r) = \sqrt{n(n+1)}\, \hat{C}_{mn} j_n(\pi\beta_{nl}r) = 0 \quad \text{on } \Gamma.$$

Thus we have

$$\int_\Omega \chi_{nml} r \cdot P_n\, dV = 0 \tag{A2}$$

for every triple (nml). Since the set of χ_{nml} is a complete orthonormal basis in $L^2(\Omega)$ [14], we have $P_n = 0$ in Ω.

Also, by assumption

$$0 = \int_\Omega P \cdot M_{nml}\, dV = \frac{1}{k^2} \int_\Omega (\nabla \times (\nabla \times P)) \cdot (\nabla \times (\psi_{nml} r\bar{a}_r)\, dV.$$

Let $\nabla \times P = R$. Note that

$$\nabla \times (\nabla \times R) = \nabla \times (k^2 P) = k^2 R.$$

Thus

$$\begin{aligned} 0 &= \frac{1}{k^2} \left\{ \int_\Omega (\psi_{nml} r\bar{a}_r) \cdot (\nabla \times (\nabla \times R))\, dV - \int_\Gamma \psi_{nml} r\bar{a}_r \cdot (\bar{a}_r \times (\nabla \times R)\, ds \right\} \\ &= \int_\Omega \psi_{nml} r\bar{a}_r \cdot R\, dV. \end{aligned}$$

Thus we have

$$\int_\Omega \psi_{nml} r R_n\, dV = 0 \tag{A3}$$

for every (nml). Since the set of ψ_{nml} forms an orthonormal basis for $L^2(\Omega)$ (14) we have $R_n = 0$ in Ω. In [18], we have Theorem A.1.

THEOREM A.1 (Lamb). *Let A be a vector field defined and of class C^1 in a domain $a<r<b$, and suppose*

$$\nabla \cdot A = 0, \qquad \bar{a}_r \cdot A = 0,$$
$$\bar{a}_r \cdot (\nabla \times A) = 0 \quad \text{in } a<r<b.$$

Then $A=0$ in $a<r<b$.

Now apply Theorem A.1, taking $A=P$, with $a=0$, $b=1$. Since $P_n=0$, and $(\nabla \times P)_n = 0$, we have $P=0$ in Ω.

Appendix B. Here, the normalization constants λ_{nl} and ρ_{nl}, defined in (4.14), (4.15), will be computed. The following identities for spherical Bessel functions [12] will be needed:

(1) $$\frac{2n+1}{z} j_n(z) = j_{n-1}(z) + j_{n+1}(z),$$

(2) $$(2n+1)\frac{d}{dz} j_n(z) = n j_{n-1}(z) - (n+1) j_{n+1}(z),$$

(3) $$\int [j_n(z)]^2 z^2 \, dz = \frac{z^3}{2}[j_n^2(z) - j_{n-1}(z) j_{n+1}(z)],$$

(4) $$n(n+1) j_n(kr) - \frac{d}{dr}\left[r^2 \frac{d}{dr} j_n(kr) \right] = k^2 r^2 j_n(kr).$$

Computation of $(\lambda_{nl})^2$. We have

$$\begin{aligned}(\lambda_{nl})^2 &= n(n+1)\int_0^1 [j_n(\pi\gamma_{nl} r)]^2 r^2 \, dr \\ &= \frac{n(n+1)}{\pi\gamma_{nl}} \int_0^{\pi\gamma_{nl}} j_n^2(z)\left(\frac{z}{\pi\gamma_{nl}}\right)^2 dz \\ &= \frac{n(n+1)}{2}[j_n^2(\pi\gamma_{nl}) - j_{n-1}(\pi\gamma_{nl}) j_{n+1}(\pi\gamma_{nl})].\end{aligned}$$

Adding (1) and (2) gives

$$j_n(z) + z j_n'(z) = \frac{z}{2n+1}[(n+1) j_{n-1}(z) - n j_{n+1}(z)].$$

From (5.1),

$$0 = (n+1) j_{n-1}(\pi\gamma_{nl}) - n j_{n+1}(\pi\gamma_{nl}).$$

Thus

$$\begin{aligned}(2n+1)^2\left\{\frac{j_n(\pi\gamma_{nl})}{\pi\gamma_{nl}}\right\}^2 &= [j_{n-1}(\pi\gamma_{nl}) + j_{n+1}(\pi\gamma_{nl})]^2 \\ &= j_{n-1}(\pi\gamma_{nl})\left[\frac{n}{n+1}\right] j_{n+1}(\pi\gamma_{nl}) + 2 j_{n-1}(\pi\gamma_{nl}) j_{n+1}(\pi\gamma_{nl}) \\ &\quad + j_{n-1}(\pi\gamma_{nl})\left[\frac{n+1}{n}\right] j_{n+1}(\pi\gamma_{nl}) \\ &= \frac{(2n+1)^2}{n(n+1)} j_{n-1}(\pi\gamma_{nl}) j_{n+1}(\pi\gamma_{nl}).\end{aligned}$$

Therefore,

$$(B1) \qquad (\lambda_{nl})^2 = \frac{n(n+1)}{2}\left[1 - \frac{n(n+1)}{(\pi\gamma_{nl})^2}\right] j_n^2(\pi\gamma_{nl}).$$

Computation of $(\rho_{nl})^2$.

$$\int_0^1 \frac{[n(n+1)j_n(\pi\beta_{nl}r)]^2 + n(n+1)\{d/dr[rj_n(\pi\beta_{nl}r)]\}^2 r^2\, dr}{(\pi\beta_{nl}r)^2} = (\rho_{nl})^2.$$

Now (omitting at times the argument of j_n)

$$\int_0^1 \frac{d}{dr}[rj_n(\pi\beta_{nl}r)] \cdot \frac{d}{dr}[rj_n(\pi\beta_{nl}r)]\, dr$$

$$(*) \qquad = rj_n(\pi\beta_{nl}r)\frac{d}{dr}[rj_n(\pi\beta_{nl}r)]\Big|_0^1 - \int_0^1 rj_n \frac{d^2}{dr^2}[rj_n]\, dr$$

$$= -\int_0^1 rj_n \frac{d^2}{dr^2}[rj_n]\, dr.$$

Since

$$rj_n \frac{d^2}{dr^2}[rj_n] = rj_n \frac{d}{dr}\left(j_n + r\frac{d}{dr}j_n\right)$$

$$= j_n\left(2r\frac{d}{dr}j_n + r^2\frac{d^2}{dr^2}j_n\right)$$

$$= j_n \frac{d}{dr}\left[r^2\frac{d}{dr}j_n\right]$$

we have that

$$(*) = -\int_0^1 j_n(\pi\beta_{nl}r) \cdot \frac{d}{dr}\left[r^2\frac{d}{dr}j_n(\pi\beta_{nl}r)\right] dr$$

$$= \int_0^1 j_n(\pi\beta_{nl}r)((\pi\beta_{nl}r)^2 j_n(\pi\beta_{nl}r) - n(n+1)j_n(\pi\beta_{nl}r))\, dr,$$

the last equality from (4). Thus,

$$(\rho_{nl})^2 = \frac{n(n+1)}{(\pi\beta_{nl})^2}\int_0^1 \{n(n+1)j_n^2(\pi\beta_{nl}r) - n(n+1)j_n^2(\pi\beta_{nl}r) + (\pi\beta_{nl}r)^2 j_n(\pi\beta_{nl}r)\}\, dr$$

$$= (\text{by } 3)\ \frac{n(n+1)}{2}[j_n^2(\pi\beta_{nl}) - j_{n-1}(\pi\beta_{nl})j_{n+1}(\pi\beta_{nl})].$$

By (1), $j_{n-1}(\pi\beta_{nl}) = -j_{n+1}(\pi\beta_{nl})$. Therefore

$$(B2) \qquad (\rho_{nl})^2 = \frac{n(n+1)}{2} j_{n+1}^2(\pi\beta_{nl}).$$

These computations are similar to those in [12] used to compute normalization constants for eigenfunctions that satisfy the "perfect conductor" boundary condition, zero tangential component and zero normal curl on Γ. Interchanging the positions of $\pi\gamma_{nl}$ and $\pi\beta_{nl}$ in our expressions for M and N gives these eigenfunctions. It is advantageous that simple expressions for the normalization constants, which resulted for the "perfect conductor" case, also result in our case.

Acknowledgment. The results in this paper come from my Ph.D. dissertation, University of Wisconsin, Madison, Wisconsin, 1986. I wish to thank my thesis advisor, Professor David Russell, for useful advice and many helpful discussions.

REFERENCES

[1] J. M. BLATT AND V. F. WEISKOPF, *Theoretical Nuclear Physics*, John Wiley, New York, 1952.

[2] E. B. BYKHOVSKII, *The solution of the mixed problem for the system of Maxwell equations in the case of an ideally conducting boundary*, Vestnik Leningrad. Univ. Math., 13 (1957), pp. 50–56; candidate's dissertation, Leningrad University, Leningrad, USSR, 1958.

[3] E. B. BYKHOVSKII AND N. V. SMIRNOV, *On the orthogonal decomposition of the space of vector functions square summable in a given domain and the operators of vector analysis*, Trudy Mat. Inst. Steklov, 59 (1960), pp. 6–36.

[4] A. EDMONDS, *Angular Momentum in Quantum Mechanics*, Princeton University Press, Princeton, NJ, 1957.

[5] K. O. FRIEDRICHS, *Mathematical methods of electromagnetic theory*, Courant Institute of Mathematical Sciences, New York University, New York, 1974.

[6] K. D. GRAHAM, *Separation of Eigenvalues of the Wave Equation for the Unit Ball in R^n*, Studies in Applied Mathematics LII, 1973, pp. 329–343.

[7] K. D. GRAHAM AND D. L. RUSSELL, *Boundary value control of the wave equation in a spherical region*, SIAM J. Control, 13 (1975), pp. 174–196.

[8] J. D. JACKSON, *Classical Electrodynamics*, 2nd ed., John Wiley, New York, 1975.

[9] O. A. LADYSHENSKAYA AND V. A. SOLONIKOV, *The linearization principle and invariant manifolds for problem of magnetohydrodynamics*, J. Soviet Math., 8 (1977), pp. 384–422. Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI), 38 (1973), pp. 46–93.

[10a] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[10b] ———, *Control of Distributed Singular Systems*, Gauthier-Villars, Paris, 1985.

[11] J. MATHEWS, *Tensor spherical harmonics*, Graphic Arts, California Institute of Technology, Pasadena, CA, 1981.

[12] P. M. MORSE AND H. FESHBACH, *Methods of Theoretical Physics*, McGraw-Hill, New York, 1953.

[13] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[14] D. L. RUSSELL, *A Unified Boundary Controllability Theory for Hyperbolic and Parabolic Partial Differential Equations*, Studies in Applied Mathematics LII, 1973, pp. 189–211.

[15] ———, *The Dirichlet-Neumann boundary control problem associated with Maxwell's equations in a cylindrical region*, SIAM J. Control Optim., 24 (1986), pp. 199–229.

[16] R. TEMAM, *Navier-Stokes Equations: Theory and Numerical Analysis*, North-Holland, New York, 1984.

[17] H. WEYL, *The method of orthogonal projection in potential theory*, Duke Math. J., 7 (1940), pp. 411–444.

[18] C. H. WILCOX, *Debye potentials*, J. Math. Mech., 6 (1957), pp. 167–201.

[19] K. YOSIDA, *Functional Analysis*, Academic Press, New York, 1965.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 320-341, March 1990

004

CONVERGENCE OF SUBOPTIMAL CONTROLS: THE POINT TARGET CASE*

H. O. FATTORINI†

Abstract. Sequences of suboptimal controls are considered for arbitrary optimal control problems in the setting of general input-output systems involving a point target condition. A *sequence maximum principle* for these sequences is obtained using Ekeland's variational principle. This sequence maximum principle and other variants are upgraded to *convergence principles* and are applied to show convergence of sequences of suboptimal controls for quasilinear distributed parameter systems, both of parabolic and hyperbolic type. Some of the results apply, without convexity assumptions, on either the systems or on the control set.

Key words. optimal control, maximum principle, approximately optimal control

AMS(MOS) subject classifications. 93E20, 93E25

1. Introduction. Convergence of suboptimal (that is, close-to-optimal) controls for general nonlinear input-output systems was considered in [9] and [10]. The optimal control problems there involve a *set target* condition, that is, trajectories are supposed to hit a *target set* Y. In this paper we continue the study of convergence of sequences of suboptimal controls for general systems (see § 2), this time for the *point target* problem, where trajectories are required to hit a point $\bar{y}$. The strategy is about the same as that in [9] and [10] and consists of three steps. In the first step, using Ekeland's variational principle, we establish a *sequence maximum principle*, that is, an independent approximate maximum principle for each of the suboptimal controls in the sequence. In the second step, we combine these separate maximum principles into a *convergence principle.* The third step consists of translating the convergence principle into actual convergence of optimal controls. Here, we work with particular classes of systems, which we did as well in [9] and [10]. The most notable difference between the treatment of set targets and that of point targets can be roughly summarized as follows. In the set target case, the passage from the weak sequence maximum principle to the convergence principle is automatic; no controllability assumptions are used, although approximate controllability (of the linearized system) plays a role in obtaining convergence of suboptimal controls. In contrast, in the point target case treated here, the convergence principle can only be deduced from the sequence maximum principle via controllability properties of the linearized system and, in fact, may not be true unless controllability is present. The situation here is roughly the same as that in the proof of the maximum principle for general systems in [8] and makes the point target problem much more demanding than the set target problem.

We note that the convergence results in this paper depend only on suboptimality of the controls in question; thus, the particular way in which these controls are constructed is irrelevant. Hence, our results justify existing computations for approximation of optimal controls (usually carried out by penalty methods [9]) rather than proposing specific computational approaches. See also [1], [2], [14].

We note also that the final convergence theorems, when they can be obtained, refer to *strong* convergence (that is, convergence in L^p norms, $1 \leqq p < \infty$). When the control set U is bounded, convergence (of suitable subsequences) in weak topologies

* Received by the editors May 18, 1987; accepted for publication (in revised form) April 24, 1989.
† University of California, Department of Mathematics, Los Angeles, California 90024. This work was supported in part by National Science Foundation grant DMS 82-00645.

follows from weak compactness, but is not very useful from a computational point of view.

Finally, it should be pointed out that the results refer not only to convergence to optimal controls whose existence has been previously established; for instance, constructive existence theorems can be obtained in certain situations without any convexity assumptions (see § 5). The same is true, of course, of the set target problems considered in [9] and [10]. Some results in this paper were announced in [11].

2. Systems. Optimal control problems. We denote by E, F arbitrary Banach spaces, although most of the results in the following sections require both to be Hilbert spaces. U is a subset of F called the *control set.* Given $k \geqq 0$, $T > 0$, we define the *control space* $W(-k, T; U)$ as the set of all (equivalence classes of) strongly measurable F-valued functions $u = u(t)$ defined in $-k \leqq t \leqq T$ such that

$$u(t) \in U \quad \text{a.e.}$$

The space $W(-k, T; U)$ is a complete metric space equipped with the distance

$$d(u, v) = \text{meas}\,\{t: u(t) \neq v(t)\} \tag{2.1}$$

called the *Ekeland distance.*

The *trajectory* or *output* space $C(0, T; E)$ consists of all E-valued continuous functions $y(t)$ defined in $0 \leqq t \leqq T$.

A *system* is, by definition, a map

$$X: W(-k, T; U) \to C(0, T; E) \tag{2.2}$$

that satisfies the following postulates: (a) *Causality.* Let $0 \leqq \bar{t} \leqq T$. Then the *trajectory*

$$y(t, u) = (Xu)(t)$$

in $0 \leqq t \leqq \bar{t}$ depends only on u in $0 \leqq t \leqq \bar{t}$. (b) *Pointwise continuity.* For $\bar{t}$ as in (a), the map

$$u(t) \to y(\bar{t}, u)$$

from $W(-k, \bar{t}; U)$ (endowed with the Ekeland distance) into E (endowed with its original norm) is continuous. (c) *Differentiability* (with respect to spike perturbations). For every $u(t) \in W(-k, \bar{t}; U)$ there exists a set $e = e(u)$ of full measure in $0 \leqq t \leqq \bar{t}$ such that, if $s \in e$, the limit

$$\xi(\bar{t}, s, u, v, u(s)) = \lim_{r \to 0+} r^{-1}(y(\bar{t}, u_{s,r,v}) - y(\bar{t}, u)) \tag{2.3}$$

exists; here $u_{s,r,v}(t)$ is the *spike perturbation* of the control $u(t)$ defined by $u_{s,r,v}(t) = v$ in $s - r < t \leqq s$, $u_{s,r,v}(t) = u(t)$ elsewhere. We call ξ the *derivative* of X.

In this general setting, systems are meant to model (among other things) input-output relationships generated by widely different *state equations* (such as ordinary or delay differential equations, or partial differential equations with distributed or boundary control). The constant k in the control space $W(-k, T; U)$ accounts for possible time delays in the control action (see [8] for systems where $k > 0$). In the examples considered in this paper controls act instantaneously, thus $k = 0$.

We shall consider below the following *optimal control problems.* Let $X: W(-k, T; U) \to C(0, T; E)$ be a system as defined above and consider a second system $X_0: W(-k, T; U) \to C(0, T; \mathbb{R})$ ($\mathbb{R}$ the real numbers) called the *cost functional* of the problem. We denote by $\xi_0(t, s, u, v, w)$ the derivative of X_0. The *augmented system* $\tilde{X}$ is defined by

$$(\tilde{X}u)(t) = ((X_0, X)u)(t) = (y_0(t, u), y(t, u)), \tag{2.4}$$

where $y_0(t, u) = (X_0 u)(t)$. We also use the notation $\tilde{y}(t, u) = (\tilde{X}u)(t)$ for the trajectories of the augmented system in $\mathbb{R} \times E$; in the same fashion, we write $\tilde{\xi}(t, s, u, v, w) = (\xi_0(t, s, u, v, w), \xi(t, s, u, v, w))$.

Let $\bar{y}$ be a point in E (called the *target*). We consider the *optimal control problem* of identifying the times $\bar{t}$ and the controls $\bar{u}$ in $W(-k, \bar{t}; U)$ such that

$$y(\bar{t}, \bar{u}) = \bar{y} \tag{2.5}$$

and

$$y_0(\bar{t}, \bar{u}) = m = \inf y_0(t, u), \tag{2.6}$$

the infimum taken over all times $t > 0$ and all controls $u(t)$ in $W(-k, t; U)$ such that $y(t, u) = \bar{y}$; we *assume* that

$$-\infty < m < \infty. \tag{2.7}$$

A sufficient condition for the first inequality (2.7) to hold is that the cost functional $y_0(t, u)$ be *bounded below* in $W(-k, T; U)$. The second inequality simply means that there exists a control u that hits the target $\bar{y}$ at some time.

A control $u(t)$ in $W(-k, \tilde{t}; U)$ is called $(\tilde{t}, \varepsilon)$-*suboptimal if*

$$\|y(\tilde{t}, u) - \bar{y}\| \leqq \varepsilon, \qquad y_0(\tilde{t}, u) \leqq m + \varepsilon. \tag{2.8}$$

The present theory of convergence of suboptimal controls, be it in the set target or point target case, is based on *weak compactness* properties of the system X; roughly speaking, what is required is that if $\{u^n\}$ is an arbitrary sequence of controls, $u^n \in W(-k, t_n; U)$ then the corresponding sequence of trajectories $\{y(t, u^n)\}$ should have a subsequence convergent (in one sense or another) to a function $y(t)$, not necessary a trajectory of the system. A similar property is required of the variations. This will be satisfied (in different ways) by the systems in § 4 and § 8.

3. The sequence maximum principle. In this and the following sections, E and F are Hilbert spaces.

THEOREM 3.1 (*The sequence maximum principle*). *Let $\{u^n\}$ be a sequence of (t_n, u^n)-suboptimal controls with $\{t_n\}$ bounded, $\varepsilon_n \to 0$. Then there exists a subsequence of $\{u^n\}$ (which we denote by the same symbol) such that*

$$t_n \to \bar{t}, \quad y_0(t_n, u^n) \to m' \leqq m, \quad y(t_n, u^n) \to \bar{y}, \tag{3.1}$$

a second sequence $\{\tilde{u}^n\}$ of controls with $\tilde{u}^n \in W(-k, t_n; U)$ such that

$$d_n(u^n, \tilde{u}^n) \to 0, \tag{3.2}$$

(d_n the Ekeland distance in $W(-k, t_n; U)$), a sequence $\{\tilde{y}^n\} = \{(\mu_n, y^n)\}$ in $\mathbb{R} \times E$ such that $\mu_n \geqq 0$, $\|(\mu_n, y^n)\| = 1$, and a set e_n of full measure in $0 \leqq s \leqq t_n$ such that

$$\begin{aligned} &\mu_n \xi_0(t_n, s, \tilde{u}^n, v, \tilde{u}^n(s)) + \langle y^n, \xi(t_n, s, \tilde{u}^n, v, \tilde{u}^n(s))\rangle \\ &\quad = \langle \tilde{y}^n, \tilde{\xi}(t_n, s, \tilde{u}^n, v, \tilde{u}^n(s))\rangle \geqq -\delta_n \to 0 \quad (v \in U, s \in e_n). \end{aligned} \tag{3.3}$$

Proof. Obviously, we can obtain the first two statements in (3.1) by passing to a subsequence; the third follows from the definition of suboptimal control. To prove (3.2) and (3.3) we consider the function

$$F_n(u) = \{(\max(0, y_0(t_n, u) - m + \varepsilon_n))^2 + \|y(t_n, u) - \bar{y}\|^2\}^{1/2} \tag{3.4}$$

in the space $W(-k, t_n; U)$ (which is complete with respect to the distance (2.1), see [4]). Obviously, $F_n(u) > 0$ (otherwise we could hit $\bar{y}$ with value $m - \varepsilon_n < m$ of y_0). Since $y_0(t_n, u^n) - m < \varepsilon_n$,

$$F_n(u^n) \leqq \{(2\varepsilon_n)^2 + (\varepsilon_n)^2\}^{1/2} = \sqrt{5}\, \varepsilon_n = c\varepsilon_n. \tag{3.5}$$

Applying Ekeland's variational principle [3], [4] we deduce the existence of a control $\tilde{u}^n \in W(-k, t_n; U)$ such that

$$F_n(\tilde{u}^n) \leqq F(u_n) = c\varepsilon_n, \tag{3.6}$$

$$d_n(u^n, \tilde{u}^n) \leqq \sqrt{c\varepsilon_n}, \tag{3.7}$$

$$F_n(w) \geqq F_n(\tilde{u}^n) - \sqrt{c\varepsilon_n}\, d_n(w, \tilde{u}^n). \tag{3.8}$$

We exploit inequality (3.8) for spike variations $w = (\tilde{u}^n)_{s,r,v}$ of $\tilde{u}^n$ using (c) in the definition of system. Note that if g is an arbitrary function, $\max(0, g)^2$ is differentiable if g is differentiable, with derivative $= (\max(0, g))g'$. We consider two alternatives. If we have $y(t_n, \tilde{u}^n) \neq \bar{y}$, then we obtain (3.3) computing the right-sided derivative at $r = 0$ of the function $F_n((u^n)_{s,r,v})$; the vector (μ_n, y^n) is given by $(\mu_n, y^n) = (\lambda_n, x^n)/\|(\lambda_n, x^n)\|$, where

$$(\lambda_n, x^n) = (\max(0, y_0(t_n, \tilde{u}^n) - m + \varepsilon_n), y(t_n, \tilde{u}^n) - y_n). \tag{3.9}$$

On the other hand, if $y(t_n, \tilde{u}^n) = \bar{y}$ we must have $y_0(t_n, \tilde{u}^n) \geqq m$, so the maximum can be omitted from the definition of $F((u^n)_{s,r,v})$ for sufficiently small r. This time we obtain (3.3) with $(\mu_n, y^n) = (1, 0)$.

As usual, a separate theorem is required for the time optimal problem, where, since $\xi_0 = 0$, (3.3) may be empty (in case $\mu_n = 0$).

THEOREM 3.2 (*The sequence maximum principle for the time optimal problem*). *Let* $\{u^n\}$ *be a sequence of* (t_n, u^n)*-suboptimal controls with* $\varepsilon_n \to 0$ *and*

$$t_n < \bar{t}. \tag{3.10}$$

Then there exists a subsequence of $\{u^n\}$ (*denoted by the same symbol*) *such that*

$$t_n \to \tilde{t} \leqq \bar{t}, \quad y(t_n, u^n) \to \bar{y}, \tag{3.11}$$

($\bar{t}$ *the optimal time*), *a second sequence* $\{\tilde{u}^n\}$ *of controls* $\tilde{u}^n \in W(-k, t_n; U)$ *such that*

$$d_n(u^n, \tilde{u}^n) \to 0, \tag{3.12}$$

a sequence $\{y^n\}$ *in* E *such that* $\|y^n\| = 1$ *and a set* e_n *of full measure in* $0 \leqq s \leqq t_n$ *such that*

$$\langle y^n, \xi(t_n, s, \tilde{u}^n, v, \tilde{u}^n(s))\rangle \geqq -\delta_n \to 0 \qquad (v \in U, s \in e_n). \tag{3.13}$$

The proof is entirely similar to that of Theorem 3.1: this time we consider the function

$$F_n(u) = \|y(t_n, u) - \bar{y}\|, \tag{3.14}$$

which satisfies $F_n(u) > 0$, since $\bar{y}$ cannot be hit in time $t_n < \bar{t}$. The vector y^n in (3.13) is $y^n = x^n/\|x^n\|$, where

$$x^n = y(t_n, \tilde{u}^n) - \bar{y}. \tag{3.15}$$

For a similar argument, see [7], [8, § 6].

Remark 3.3. Condition (3.10) (which is indispensable to guarantee positivity of the function (3.14), required for differentiation of the norm) is somewhat inconvenient. In some cases, it can be circumvented by simply replacing the sequence $\{t_n\}$ by a second sequence $\{\tilde{t}_n\}$, with $\tilde{t}_n < \bar{t}$ and

$$\|y(\tilde{t}_n, u^n) - y(t_n, u^n)\| \to 0 \quad \text{as } n \to \infty. \tag{3.16}$$

In fact, if (3.16) holds, the (t_n, ε_n)-suboptimal sequence $\{u^n\}$ will be as well $(\tilde{t}_n, \varepsilon_n)$-suboptimal, and Theorem 3.2 applies. Since $t_n \to \bar{t}$, (3.16) will hold if the set $\{y(t, u)\}$

of all trajectories of X corresponding to all $u \in W(-k, T; U)$ is *equicontinuous.* This is in fact the case, for instance, when X is defined by a finite-dimensional differential system. The sequence maximum principle has to be modified in obvious ways; in (3.12) d_n is the Ekeland distance in $W(-k, \tilde{t}_n; U)$, and (3.13) becomes

$$\langle y^n, \xi(\tilde{t}_n, s, \tilde{u}^n, v, \tilde{u}_n(s))\rangle \geqq -\delta_n \to 0 \qquad (v \in U, s \in e_n). \tag{3.17}$$

4. Systems described by abstract differential equations. Generally speaking, a sequence maximum principle becomes a *convergence principle* if we can show that the sequence $\{(\mu_n, y^n)\}$ in (3.3) has a subsequence (denoted with the same symbol) such that

$$(\mu_n, y^n) \to (\mu, y) \neq 0. \tag{4.1}$$

The convergence principle is *strong* (respectively, *weak*) if the convergence in (4.1) is strong (respectively, weak); note that, since $\|(\mu_n, y^n)\| = 1$, in the first case, the requirement that $(\mu, y) \neq 0$ is unnecessary.

This definition is modified in an obvious way for the time optimal problem, where the sequence maximum principle is (3.13); here, (4.1) becomes

$$y^n \to y \neq 0. \tag{4.2}$$

In some situations (as seen below), the weak convergence principle implies *strong* convergence of sequences of suboptimal controls, usually via compactness properties of the system. The strong convergence principle implies strong convergence of sequences of suboptimal controls without any added compactness properties, but it is much more difficult to prove.

Of course, in case E is finite-dimensional, the passage from the sequence maximum principle to the strong convergence principle is automatic via the Bolzano–Weierstrass theorem.

We consider below systems defined by a semilinear initial value problem

$$y'(t) = Ay(t) + f(t, y(t), u(t)) \qquad (0 \leqq t \leqq T), \tag{4.3}$$

$$y(0) = y^0, \tag{4.4}$$

where A is the infinitesimal generator of a strongly continuous semigroup $\{S(t): t \geqq 0\}$ in the Banach space E. By definition, a *solution* of (4.3)–(4.4) is a continuous solution of the integrated version

$$y(t) = S(t)y^0 + \int_0^t S(t-\sigma) f(\sigma, y(\sigma), u(\sigma))\, d\sigma \qquad (0 \leqq t \leqq T). \tag{4.5}$$

We assume that U is bounded and that $f(t, y, u)$ satisfies the following assumptions:

(F) $f(t, y, u)$ has a Fréchet derivative $\partial_y f(t, y, u)$ with respect to y and f (respectively, $\partial_y f$) is continuous (respectively, strongly continuous) and bounded on bounded subsets of $[0, T] \times E \times U$.

Under these hypotheses, (4.5) can be solved as usual by successive approximations: setting $y^0(t, u) = S(t)y^0$,

$$y^{m+1}(t, u) = S(t)y^0 + \int_0^t S(t-\sigma) f(\sigma, y^m(\sigma, u)), u(\sigma))\, d\sigma \qquad (0 \leqq t \leqq T). \tag{4.6}$$

Since, in general, the approximations $y^m(t, u)$ in (4.6) will not be convergent in the whole interval $[0, T]$, we shall need to construct the solution "in pieces," that is, first in an interval $[0, T_1]$, then in an interval $[T_1, T_2], \cdots$ etc., so that, given T_k in the

interval $[0, T)$ we shall have to examine as well the sequence of successive approximations

$$y^{m+1}(t, u) = S(t-T_k)y(T_k, u) + \int_{T_k}^{t} S(t-\sigma)f(\sigma, y^m(\sigma, u), u(\sigma))\, d\sigma \tag{4.7}$$

in $t \geqq T_k$. Obviously, in order to extend the solution to the whole interval $[0, T]$ it suffices to show that the sequence $\{y^m(t, u)\}$ is convergent in an interval $[T_k, T_{k+1}]$, where $T_{k+1} - T_k$ is bounded away from zero uniformly in k. To insure this, it is enough to have an a priori bound

$$\|y(t)\| \leqq C \qquad (0 \leqq t \leqq T_0) \tag{4.8}$$

for any solution of (4.5) in an arbitrary interval $[0, T_0]$, where C does not depend on T_0 or on $u \in W(0, T; U)$. In fact, assume that (4.8) holds. Let M be a bound for $\|S(t)\|$ in $[0, T]$ and let K be such that

$$\|f(t, y, u)\| \leqq K \qquad (0 \leqq t \leqq T,\ \|y\| \leqq MC+1,\ u \in U).$$

Then we obtain from (4.8) that, if $\|y^m(t, u)\| \leqq MC+1$ in $[T_k, T_{k+1}]$ (which is obviously true for $y_0(t, u) = y(T_k, u)$), we shall have

$$\|y^{m+1}(t, u)\| \leqq MC + MK(t-t_k) \leqq MC+1,$$

if $T_{k+1} - T_k \leqq 1/MK$. Since the approximations defined by (4.8) are bounded in $t_k \leqq t \leqq t_{k+1}$, (by $MC+1$), it is clear that we can estimate $\|y^{m+1}(t, u) - y^m(t, u)\|$ in terms of $\|y^m(t, u) - y^{m-1}(t, u)\|$ using the Lipschitz constant L for $f(t, y, u)$ in $\|y\| \leqq MC+1$, thus the sequence $\{y^m(t, u)\}$ is absolutely and uniformly convergent in $T_k \leqq t \leqq T_{k+1}$ if, in addition, $T_{k+1} - T_k < 1/L$.

The way to establish (4.8) depends on the particular equation under study; see [9] for quasilinear parabolic equations and [10] for the quasilinear hyperbolic case.

Under (4.8) and the rest of the assumptions on $f(t, y, u)$ and A we can prove that the map defined by

$$(Xu)(t) = y(t, u) = y(t), \tag{4.9}$$

where $y(t)$ is the only solution of (4.5), satisfies postulates (a), (b), and (c) in § 2; the derivative $\xi(t, s, u, v, w)$ in (2.3) is

$$\xi(t, s, u, v, w) = S(t, s; u)\{f(s, y(s, u), v) - f(s, y(s, u), w)\} \tag{4.10}$$

where $S(t, s; u)$ is the solution operator of the *linearized equation* $z'(t) = (A + B(t))z(t)$, $B(t; u) = \partial_y f(t, y(t, u), u)$, that is, the only strongly continuous solution of the operator equation

$$S(t, s; u)z = S(t-s)z + \int_s^t S(t-\sigma)B(\sigma; u)S(\sigma, s; u)z\, d\sigma \qquad (0 \leqq s \leqq t \leqq T). \tag{4.11}$$

For complete proofs, see [8]. In particular, we point out that, for u fixed, $S(t, s; u)$ is strongly continuous in its domain of definition (we shall need later in §§ 5 and 6 information on the u-dependence of $S(t, s; u)$). If $S(t)$ is a group, $S(t, s; u)$ of course exists and is strongly continuous in $0 \leqq s, t \leqq T$.

5. Convergence principles and strong convergence of suboptimal controls. We show in this section how convergence principles can be used to establish L^p-convergence of sequences of suboptimal controls $(1 < p < \infty)$ for the system defined by the initial value problem (4.3)-(4.4). We assume from now on that the control set U is bounded and closed and that the nonlinear term $f(t, y, u)$ in (4.3) is of the form

$$f(t, y, u) = f(t, y) + Bu, \tag{5.1}$$

where $B: F \to E$ is a bounded operator. Finally, we assume that either the semigroup $S(t)$ is compact for all $t > 0$ or that B is compact.

The first manipulations do not use compactness of $S(t)$. Noting the computation of the function $\xi(t, s, u, v, w)$ in (4.9), we can write the sequence maximum principle (3.3) in the form

$$\begin{aligned} &\mu_n \xi_0(t_n, s, \tilde{u}^n, v, \tilde{u}^n(s)) \\ &\quad + \langle y^n, S(t_n, s; \tilde{u}^n) B(v - \tilde{u}^n(s)) \rangle \geqq -\delta_n \to 0 \qquad (v \in U, s \in e). \end{aligned} \tag{5.2}$$

The sequence maximum principle (3.13) for the time optimal problem is

$$\langle y^n, S(t_n, s; \tilde{u}^n) B(v - \tilde{u}^n(s)) \rangle \geqq -\delta_n \to 0 \qquad (v \in U, s \in e). \tag{5.3}$$

LEMMA 5.1. *Assume that either B is compact or that $S(t)$ is compact for $t > 0$. Then the operator*

$$(\Pi u)(t) = \int_0^t S(t - \sigma) B u(\sigma)\, d\sigma \tag{5.4}$$

from $L^2(0, T; F)$ into $C(0, T; E)$ (the last space endowed with its usual supremum norm) is compact.

Lemma 5.1 for $S(t)$ compact is Lemma 6.1 in [9]; the proof where B is compact instead is essentially the same, since the kernel $S(t)B$ in the integral (5.4) is compact.

In what follows, we shall consider convergence of sequences $\{u^n\}$ where each control u^n lives in a different space $W(-k, t_n; U)$. Assume that $t_n \to t_0$. We say that $\{u^n\}$ *converges weakly* to $u \in W(-k, t_n; U)$ if u^n, extended to $t \leqq t_n$ (if $t_n < t_0$) by setting $u^n(t) = 0$ there, or chopped off at t_0 (if $t_n > t_0$) converges weakly in $L^2(0, t_0; U)$. A similar meaning will be given to other types of convergence.

We write below

$$V = \overline{\text{conv}}\,(U), \tag{5.5}$$

where $\overline{\text{conv}}$ denotes closed convex hull.

THEOREM 5.2. *Under the same hypotheses of Lemma 5.1, let $\{u^n\}$ be a sequence of controls such that $u^n \in W(0, t_n; U)$. Assume that $t_n \to t_0$. Then there exists a subsequence of $\{u^n\}$ (which we denote by the same symbol) and a $\bar{u} \in W(0, t_0; V)$ such that*

$$u^n(t) \to \bar{u}(t) \quad \textit{weakly in } L^2(0, t_0; F), \tag{5.6}$$

$$y(t, u^n) \to y(t, \bar{u}) \quad \textit{in } C(0, t_0; E). \tag{5.7}$$

Proof. Let $\{u^n\}$ be the sequence in the statement of Theorem 5.2, and let $T > t_n$; extend each u^n to $L^2(0, T; F)$ by setting $u^n(t) = 0$ in $t > t_n$. We may assume, by passing to a subsequence, that $\{u^n\}$ is weakly convergent, so that (5.6) will hold. Since $W(0, T; V)$ is bounded, closed and convex (hence weakly closed) in $L^2(0, T; F)$, the limit $\bar{u}$ belongs to $W(0, T; V)$ as well. We take a look at the approximations $y^m(t, u^n)$, $m = 1, 2, \cdots$ used to compute $y(t, u^n)$ (see (4.6)): these are $y^0(t, u^n) = 0$,

$$\begin{aligned} y^{m+1}(t, u^n) = S(t) y_0 &+ \int_0^t S(t - \sigma) f(\sigma, y^m(\sigma, u^n))\, d\sigma \\ &+ \int_0^t S(t - \sigma) B u^n(\sigma)\, d\sigma. \end{aligned} \tag{5.8}$$

Applying Lemma 5.1 inductively, we deduce that

$$y^m(t, u^n) \to y^m(t, \bar{u}) \tag{5.9}$$

uniformly in $0 \leqq t \leqq T$ as $n \to \infty$, where the $y^m(t, \bar{u})$ are the approximations to $y(t, \bar{u})$ defined by (4.6). On account of the fact that both $\{y^m(t, u)\}$ and $\{y^m(t, u^n)\}$ converge absolutely and uniformly as $m \to \infty$ (independently of n) in an interval $0 \leqq t \leqq T_1$ (see the observations after (4.6)), (5.9) shows that

$$y^m(t, u^n) \to y(t, \bar{u}) \tag{5.10}$$

uniformly in $0 \leqq t \leqq T_1$. Applying the same argument in the intervals $[T_1, T_2]$, $[T_2, T_3], \cdots$ referred to after (4.7) (the lengths of these intervals do not tend to zero in view of the a priori estimate (4.8)), we deduce that the convergence in (5.10) is uniform in the interval $0 \leqq t \leqq T$. This ends the proof of Theorem 5.2.

Remark 5.3. We note that the conclusion of Theorem 5.2 can be made more precise: if $\{u^n\}$ satisfies (5.6), then the same sequence $\{u^n\}$ will satisfy (5.7), without need of passing to a subsequence. The proof uses a standard trick: if the full sequence $\{u^n\}$ fails to satisfy (5.7) there exists a subsequence which stays at a positive distance from $y(t, \bar{u})$. Applying the argument in Theorem 5.2 to this subsequence, a contradiction ensues.

In the following two results, we assume that $S(t)$ is compact for all $t > 0$; recall that this implies that $S(t)$ is continuous in $t > 0$ in the uniform topology of operators.

LEMMA 5.4. *Let $h > 0$. The (E)-valued function*

$$(t, s, u) \to S(t, s; u) \tag{5.11}$$

is uniformly continuous (in the uniform topology of operators) in $0 \leqq s \leqq t - h \leqq T$, $u \in W(0, T; U)$ (u measured in the weak $L^2(0, T; E)$-topology).

LEMMA 5.5. *Let $0 \leqq s < t \leqq T$, $u \in W(0, t; U)$. Then the operator $S(t, s; u): E \to E$ is compact.*

For proofs of Lemmas 5.3 and 5.4, see [9, Thms. 6.3 and 6.4].

We show below how the weak convergence principle can be used to show strong convergence of sequences of suboptimal controls. We note that in their present form, both the weak and the strong convergence principles have been proved only in finite-dimensional spaces. However, slightly modified versions of the sequence maximum principle and of the convergence principles hold in infinite-dimensional spaces (see §§ 6 and 7) and the arguments showing that the convergence principle implies convergence of suboptimal controls are essentially the same, thus we take no advantage below of the finite dimensionality of the space.

All assumptions are in force: the control set U is bounded and closed and the nonlinear term is assumed to be of the form (5.1), with $f(t, y)$ satisfying assumption (F) in § 4; we also suppose that either B is compact or that $S(t)$ is compact for $t > 0$. For simplicity, we consider only the time optimal problem.

Assume that the weak convergence principle (5.3)-(4.2) holds for the sequence $\{\tilde{u}^n\}$ associated by Theorem 3.2 with a sequence $\{u^n\}$ of suboptimal controls. Write

$$B^*S(t_n, s; \tilde{u}^n)^*y^n = (B^*S(t_n, s; \tilde{u}^n)^* - B^*S(\bar{t}, s; \bar{u})^*)y^n + B^*S(\bar{t}, s; \bar{u})^*y^n. \tag{5.12}$$

Since each $B^*S(t, s; \bar{u})^*$ is compact we may assume, passing if necessary to a subsequence, that $B^*S(\bar{t}, s; \bar{u})^*y^n \to B^*S(\bar{t}, s; \bar{u})^*y$ strongly, y the limit in (4.2); on the other hand, using Lemma 5.4 in case $S(t)$ is compact or compactness of B^* when B is compact we deduce that $(B^*S(t_n, s; \tilde{u}^n)^* - B^*S(\bar{t}, s; \bar{u})^*)y^n \to 0$ strongly. In any case,

$$z_n(t) = B^*S(t_n, s; \tilde{u}^n)^*y^n \to z(t) = B^*S(\bar{t}, s; \bar{u})^*y \text{ strongly} \tag{5.13}$$

and we obtain from (5.3) that

$$\langle z(t), v - \tilde{u}^n(s)\rangle \geqq -\delta_n \to 0 \qquad (v \in U, s \in e_n). \tag{5.14}$$

For $z \neq 0$ let $U(z, \delta)$ be the set of all $u \in U$ such that

$$\langle z, v-u\rangle \geqq -\delta \quad \text{for all } v \in U \tag{5.15}$$

(for the geometry of the situation see [9]). Then if the control set U satisfies

$$\operatorname{diam} U(z, \delta) \to 0 \quad \text{as } \delta \to 0 \tag{5.16}$$

for all z with $z \neq 0$, it follows from (5.14) that the sequence $\{\tilde{u}^n(t)\}$ will be pointwise convergent in the set d where

$$z(t) = B^* S(\bar{t}, s; \bar{u})^* y \neq 0. \tag{5.17}$$

Since U is bounded, pointwise convergence in d is equivalent to $L^p(d)$-convergence for any $1 < p < \infty$.

Naturally, no convergence results are obtained outside of d, thus it is important to have conditions that guarantee that d has full measure in $0 \leqq t \leqq \bar{t}$. Conditions to that effect appear hard to come by for general quasilinear equations. In the linear case we have $S(t, s; u) = S(t-s)$ ($S(t)$ the semigroup generated by A). If $S(t)$ is analytic, that d be of full measure in $0 \leqq t \leqq \bar{t}$ is equivalent to *approximate controllability* of the system, one of whose formulations is

(C) if $B^* S(\bar{t}, s; \bar{u})^* y = 0$ in a set of positive measure then $y = 0$.

We note finally that the convergence conclusions are for the auxiliary sequence $\{\tilde{u}^n\}$ rather than for the original sequence $\{u^n\}$ of suboptimal controls. However, this makes no difference since by virtue of (3.2) and the definition of the distance d,

$$\operatorname{meas}\{t: u^n(t) \neq \tilde{u}^n(t)\} \to 0. \tag{5.18}$$

It is easily seen ([9]) that assumption (5.16) implies convexity of U; in fact, (5.16) is equivalent to strict convexity when F is finite-dimensional. However, convergence results can be obtained even in the nonconvex case. Call the vector $z \neq 0$ (that is, the direction z) *improper* with respect to the control set U if (5.16) fails to hold for x, that is, if

$$\limsup_{\delta \to 0} \operatorname{diam} U(z, \delta) > 0. \tag{5.19}$$

Assume that the set of improper directions is finite or at most countable. Then the conclusions will still hold if we can show that the vector $z(t)$ cannot be parallel to any direction except in a null set. As an example, consider the linear case $f = 0$, with $S(t)$ analytic and

$$B = I \tag{5.20}$$

(which insures that the approximate controllability condition (C) is satisfied) and that there exists a vector $z \neq 0$ such that

$$S(\bar{t}-s)^* y = \eta(s) z \tag{5.21}$$

in a set e of positive measure in $0 \leqq s \leqq \bar{t}$. (It is enough to assume that e is an infinite set.) Equality (5.21) means that $S(\bar{t}-s)^* y$ belongs to the one-dimensional subspace $H(z)$ generated by z, thus it follows from analyticity that $S(\bar{t}-s)^* y$ belongs to $H(z)$ for all $s < \bar{t}$. Hence, (5.21) actually holds for all $s < \bar{t}$. If $\eta(\bar{t}) = 0$, $y = 0$, which is impossible. Accordingly, it follows that $z = \eta(\bar{t})^{-1} y \in D(A)$ and

$$A^* S(\bar{t}-s)^* y = -\eta'(s) z = -\eta'(s)/\eta(\bar{t}) y. \tag{5.22}$$

Setting $s = \bar{t}$ we see that y (or, equivalently, z) is an eigenvector of A corresponding to the eigenvalue $-\eta'(\bar{t})/\eta(\bar{t})$. Accordingly, if $y_1, y_2, \cdots$ are the eigenvectors of A corresponding to real eigenvalues, we obtain conversions without the assumption that U is convex; we require instead that none of the y_k be an improper direction with respect to U. This requires knowledge of all the eigenvalues of A and is thus difficult to verify, except when A has no real eigenvalues. We do not know of a characterization of sets U having a finite or countable set of improper directions even in dimension 2. It is possible that a set given by a polar equation

$$\{(r, \theta);\ 0 \leqq r \leqq r(\theta),\ 0 \leqq \theta \leqq 2\pi\}$$

where $r(\theta)$ is sufficiently smooth has that property. That U has a finite set of improper directions can be verified for many choices of $r(\theta)$ (say, functions having a finite number of local maxima).

Remark 5.6. Some obvious simplifications are available in the finite-dimensional case. We may assume that $A = 0$; a condition that guarantees (4.8) is

$$\langle y, f(t, y, u)\rangle \leqq (1 + \|y\|^2) \quad (t \in \mathbb{R},\ y \in E,\ u \in F).$$

It can be easily shown that the set $\{y(t, u)\}$ of all trajectories corresponding to all $u \in W(0, T; U)$ is equi-Lipschitz continuous.

6. The convergence principle: abstract parabolic equations. We show here that for a certain class of quasilinear abstract parabolic equations, a modified convergence principle can be proved that implies L^p-convergence $(1 < p < \infty)$ of sequences of suboptimal controls. However, the definition of suboptimal control will have to be modified as well.

We consider again the system (4.9) defined by the initial value problem (4.3)-(4.4). Translating A if necessary we may assume that

$$\|S(t)\| \leqq C e^{-\beta t} \qquad (t \geqq 0) \tag{6.1}$$

so that fractional powers $(-A)^{\alpha}$ exist for all α and are bounded for $\alpha < 0$. (In particular, A^{-1} is a bounded operator.) We require $S(t)$ to be an analytic semigroup, so that for every $\alpha > 0$ we have a bound of the form

$$\|(-A)^{\alpha} S(t)\| \leqq C t^{-\alpha} e^{-\beta t} \qquad (t > 0). \tag{6.2}$$

We assume the nonlinear term $f(t, y, u)$ is of the form (5.1). Assumptions (F) on $f(t, y)$ will have to be strengthened:

(F_α) For some $\alpha > 0$ $(-A)^{\alpha} f(t, y)$ has a Fréchet derivative $\partial_y(-A)^{\alpha} f(t, y)$ with respect to y and $(-A)^{\alpha} f(t, y)$ (respectively, $\partial_y(-A)^{\alpha} f(t, y)$) is continuous (respectively, strongly continuous) and bounded on bounded subsets of $[0, T] \times E \times U$.

Since $(-A)^{-\alpha}$ is a bounded operator, it is clear that $f(t, y)$ satisfies (F); in particular

$$\partial_y(-A)^{\alpha} f(t, y) = (-A)^{\alpha} \partial_y f(t, y). \tag{6.3}$$

We make some changes in the theory in § 5, beginning with the definition of suboptimal controls. Approximation to the target will have to be measured in a stronger norm, namely the graph norm of A. We *assume* now that the target point $\bar{y}$ belongs to $D(A)$, the domain of A. A control $u(t)$ is $(\tilde{t}, \varepsilon)$-*suboptimal* (or, more precisely, $(A, \tilde{t}, \varepsilon)$-suboptimal) if

$$\|Ay(\tilde{t}, u) - A\bar{y}\| \leqq \varepsilon, \qquad y_0(\tilde{t}, u) \leqq m + \varepsilon. \tag{6.4}$$

The graph norm of A in (6.4) has a simple interpretation when A is a second-order partial differential operator

$$Au(x)=\sum_{j=1}^{m}\sum_{k=1}^{m}D^j(a_{jk}(x)D^k u(x))+\sum_{j=1}^{m}b_j(x)D^j(x)+c(x)u(x) \tag{6.5}$$

in a bounded domain Ω of m-dimensional Euclidean space $\mathbb{R}^m$, defined on functions that satisfy a boundary condition, for instance

$$u(x)=0 \tag{6.6}$$

on the boundary Γ of Ω. Under minimal assumptions on Ω and on the coefficients of $a_{jk}(x)$, $b_j(x)$ and $c(x)$ (see [13] for details), $D(A)$ is just the Sobolev space $H^2(\Omega)_0$ (the subspace of $H^2(\Omega)$ defined by the boundary condition (6.6)) and the graph norm is equivalent to the $H^2(\Omega)$ norm. Thus, the first inequality (6.4) means only approximation in the $H^2(\Omega)$ norm and satisfaction of the boundary condition. The same statement holds for other boundary conditions (see again [13]).

Given a sequence of (A, t_n, ε_n)-suboptimal controls we use instead of (3.4) the function

$$G_n(u)=\{(\max(0, y_0(t_n, u)-m+\varepsilon_n))^2+\|Ay(t_n, u)-A\bar{y}\|^2\}^{1/2} \tag{6.7}$$

in the space $W(0, t_n; U)$; we define $F_n(u)=+\infty$ if $y(t_n, u)$ does not belong to $D(A)$. It is easily shown that the function $F_n(u)$ is lower semicontinuous ([8, Lemma 7.4]), thus Ekeland's theorem can be applied. We obtain in this way the following version of Theorem 3.1.

THEOREM 6.1 (*The modified sequence maximum principle*). *Let* $\{u^n\}$ *be a sequence of* (A, t_n, u^n)*-suboptimal controls with* $\{t_n\}$ *bounded,* $\varepsilon_n\to 0$. *Then there exists a subsequence of* $\{u^n\}$ (*which we denote by the same symbol*) *such that*

$$t_n\to\bar{t}, \qquad y_0(t_n, u^n)\to m'\leqq m, \qquad Ay(t_n, u^n)\to A\bar{y}, \tag{6.8}$$

a second sequence $\{\tilde{u}^n\}$ *of controls with* $\tilde{u}^n\in W(-k, t_n; U)$ *such that*

$$F_n(\tilde{u}^n)<F_n(u^n)<\infty, \tag{6.9}$$

$$d_n(u^n, \tilde{u}^n)\to 0, \tag{6.10}$$

a sequence $\{\tilde{y}^n\}=\{(\mu_n, y^n)\}$ *in* $\mathbb{R}\times D(A)$ *such that* $(\mu_n)^2+\|Ay^n\|^2=1$ *and a set* e_n *of full measure in* $0\leqq s\leqq t_n$ *such that*

$$\begin{aligned}&\mu_n\xi_0(t_n, s, \tilde{u}^n, v, \tilde{u}^n(s))+\langle Ay^n, AS(t_n, s; u)B(v-\tilde{u}^n(s))\rangle\\ &\qquad\geqq -\delta_n\to 0 \quad (v\in U, s\in e_n).\end{aligned} \tag{6.11}$$

An essential ingredient in the proof is the following lemma.

LEMMA 6.2. *Let* $u\in W(0, \bar{t}; u)$, $s<\bar{t}$. *Then* (a) $S(\bar{t}, s; u)E\subseteq D(A)$, (b) *if* $r<\bar{t}-s$ *then* $y(\bar{t}, u_{s,r,v})-y(\bar{t}, u)\in D(A)$, *and*

$$AS(\bar{t}, s; u)(v-u(s))=\lim_{\rho\to 0+}r^{-1}A(y(\bar{t}, u_{s,r,v})-y(\bar{t}, u)) \tag{6.12}$$

in a set e *of full measure in* $0\leqq s<\bar{t}$.

The proof is in [8, § 7]. It is based on the integral equation

$$y(t, u)=S(t)y_0+\int_0^t S(t-\sigma)f(\sigma, y(\sigma, u))\,d\sigma+\int_0^t S(t-\sigma)Bu(\sigma)\,d\sigma \tag{6.13}$$

defining $y(t, u)$; to exploit the assumptions on f, we write the first integrand in the form

$$S(t-\sigma)f(\sigma, y(\sigma, u))=(-A)^{-\alpha}S(t-\sigma)(-A)^{\alpha}f(\sigma, y(\sigma, u)), \tag{6.14}$$

and use the estimate (6.2). We proceed in the same way in the integral equation (4.11) defining $S(t, s; u)$; here the integrand is written in the form

$$\begin{aligned} &S(t-\sigma)\partial_y f(\sigma, y(\sigma, u))S(\sigma, s; u) \\ &\quad = (-A)^{-\alpha}S(t-\sigma)(-A)^{\alpha}\partial_y f(\sigma, y(\sigma, u))S(\sigma, s; u). \end{aligned} \tag{6.15}$$

Once (6.11) has been established we take $v = v(s)$ and integrate in $0 \leqq s \leqq \bar{t}$, assuming that

$$s \to \xi_0(t_n, s, \tilde{u}^n, v(s), \tilde{u}^n(s)) \tag{6.16}$$

is integrable for any $v \in W(0, T; U)$. The result is

$$\begin{aligned} &\mu_n \int_0^{t_n} \xi_0(t_n, s, \tilde{u}^n, v(s), \tilde{u}^n(s))\, ds \\ &\quad + \left\langle Ay_n, A\int_0^{t_n} S(t_n, s; \tilde{u}^n)Bv(s)\, ds - A\int_0^{t_n} S(t_n, s; \tilde{u}^n)B\tilde{u}^n(s)\, ds \right\rangle \\ &\geqq -\delta_n \to 0. \end{aligned} \tag{6.17}$$

However, (6.17) needs some clarification. We must show first that

$$\int_0^{t_n} S(t_n, s; \tilde{u}^n)B\tilde{u}^n(s)\, ds \in D(A). \tag{6.18}$$

To see this, observe that it follows from (6.7) that

$$y(t_n, \tilde{u}^n) \in D(A). \tag{6.19}$$

We use then the integral equation (6.13) and (6.14) to deduce that

$$\int_0^{t_n} S(t-\sigma)B\tilde{u}^n(\sigma)\, d\sigma \in D(A). \tag{6.20}$$

Finally, we exploit the integral equation (4.11) defining $S(t, s; u)$, write the integrand in the form (6.15) and use the assumptions on $f(t, s, u)$ at the beginning of the section.

On the other hand, it is not necessarily true that

$$\int_0^{t_n} S(t_n, s; \tilde{u}^n)Bv(s)\, ds \in D(A), \tag{6.21}$$

thus we only claim (6.17) for those $v \in W(0, t_n; U)$ that satisfy (6.21).

We bring into play some definitions and results of [8]. Let Δ be a set in a Hilbert space H. We say that Δ has *finite codimension* if and only if there exists a closed subspace $K \subseteq H$ of finite codimension (that is, with a finite-dimensional orthogonal complement) such that

$$\Delta_K = \Pi(\overline{\text{conv}}\,(\Delta)) \tag{6.22}$$

has nonempty interior in K, where Π denotes the orthogonal projection from H onto K. A set with nonempty interior (or, more generally such that its closed convex hull has nonempty interior) has finite codimension (take $K = H$). In a finite-dimensional space, any nonempty set has finite codimension, as we see taking $K = \{0\}$.

The definition extends to families of sets. We say that a sequence $\{\Delta_n\}$ *has finite codimension* if there exists a closed subspace $K \subseteq H$ with finite codimension in H such that

$$\Delta_K = \bigcap_{n \geqq 1} \Pi(\overline{\text{conv}}\,(\Delta_n)) \tag{6.23}$$

has nonempty interior. The sequence $\{\Delta_n\}$ is called *bounded* if $\Delta_n \subseteq \Delta$ for all n, where Δ is a bounded set in H.

LEMMA 6.3. *Let $\{\Delta_n\}$ be a bounded sequence of sets in H having finite codimension. Let $\{y_n\}$ be a sequence in H such that*

$$0 < c \leqq \|y_n\| \leqq C. \tag{6.24}$$

Assume that

$$\langle y_n, z\rangle \geqq -\varepsilon_n \to 0 \quad (z \in \Delta_n,\ n = 1, 2, \cdots). \tag{6.25}$$

Then there exists a subsequence of $\{y_n\}$ that converges weakly to an element $y \in H$, $y \neq 0$.

For a proof, see [8, Lemma 5.6].

We denote below by $\Lambda(t)$ the nonlinear operator

$$\Lambda(t)u = \int_0^t S(t, s; u)Bu(s)\, ds \tag{6.26}$$

defined in $W(0, t; U)$.

LEMMA 6.4. *Let $\{\tilde{u}^n\}$ be the sequence in Theorem 6.1. Then $\Lambda(t_n)\tilde{u}^n \in D(A)$ for all n and the sequence $\{A\Lambda(t_n)\tilde{u}^n\}$ is strongly convergent in E.*

Proof. Premultiplying (4.11) by A we obtain

$$\begin{aligned}(-A)S(t, s; u)z &= (-A)S(t-s)z \\ &\quad + \int_s^t (-A)^{1-\alpha} S(t-\sigma)(-A)^{\alpha}\partial_y f(t, y(t, u))S(\sigma, s; u)z\, d\sigma.\end{aligned} \tag{6.27}$$

It follows from (6.27) and (6.2) that

$$\|AS(t, s; u) - AS(t-s)\| \leqq C(t-s)^{\alpha} \tag{6.28}$$

in $0 < s < t < T$. Write now (4.3) in the form

$$\begin{aligned}y'(t, u) &= \{A + \partial_y f(t, y(t, u))\}y(t, u) \\ &\quad + \{f(t, y(t, u)) - \partial_y f(t, y(t, u))y(t, u)\} + Bu(t)\end{aligned} \tag{6.29}$$

and express the solution $y(t, u)$ of (4.3)-(4.4) using the solution operator $S(t, s; u)$ of the linearized equation: the result is

$$\begin{aligned}y(t, u) &= S(t, 0; u)y^0 + \int_0^t S(t, \sigma; u)\{f(\sigma, y(\sigma, u)) - \partial_y f(\sigma, y(\sigma, u))y(\sigma, u)\}\, d\sigma \\ &\quad + \int_0^t S(t, \sigma; u)Bu(\sigma)\, d\sigma.\end{aligned} \tag{6.30}$$

On the left-hand side we use the fact that (6.4) implies

$$Ay(t_n, \tilde{u}^n) \to A\bar{y}. \tag{6.31}$$

On the right-hand side, we use (6.28) to replace (modulo a bounded operator) $AS(t, s; u)$ by $AS(t-s)$ in (6.30), and then write the resulting integrand in the form

$$\begin{aligned}&AS(t-\sigma)\{f(\sigma, y(\sigma, u)) - \partial_y f(\sigma, y(\sigma, u))y(\sigma, u)\} \\ &\quad = (-A)^{1-\alpha} S(t-\sigma)\{(-A)^{\alpha} f(\sigma, u)) - (-A)^{\alpha}\partial_y f(\sigma, y(\sigma, u))y(\sigma, u)\}\end{aligned} \tag{6.32}$$

and use the hypotheses (F_α) on $f(t, y)$ and Theorem 5.2. This ends the proof.

THEOREM 6.5. *Let the semigroup $S(t)$ generated by A be analytic, and let the nonlinearity $f(t, y)$ be of the form* (5.1) *with $B = I$ and $f(t, y)$ satisfying assumptions*

(F_α). *Let the control set U be closed and bounded and have 0 as an interior point. Assume, finally, that* (6.16) *is integrable for any $v \in W(0, T; U)$ and that the set of all elements of the form*

$$\int_0^t \xi_0(t, s, u, v(s), u(s))\, ds \tag{6.33}$$

with $u = \tilde{u}^n$ and $v \in W(0, t; U)$ is bounded, uniformly with respect to n. Then the sequence $\{(\mu_n, y^n)\}$ in Theorem 6.1 *has a subsequence weakly convergent in $\mathbb{R} \times D(A)$ to $(\mu, y) \neq 0$.*

Proof. Given $\bar{t} > 0$ and $u \in W(0, \bar{t}; U)$ we define (following [8]) the set $K(0, \bar{t}, u; U) \subseteq E$ as the set of all elements $z \in E$ of the form

$$z = \Lambda(t)u = \int_0^t S(t, s; u)Bv(s)\, ds. \tag{6.34}$$

The set $\tilde{K}(0, \bar{t}, u; U)$ consists of all elements of the form $(\eta, z) \in \mathbb{R} \times E$ with z given by (6.34) and η given by (6.33).

Making use of the integral equation (4.11) and the assumptions on $S(t)$ and $f(t, y)$, we show that if $y \in D(A)$ then $s \to S(t, s; u)y$ is continuously differentiable in $0 \leqq s < t$ and satisfies

$$(S(t, s; u)y)' = -S(t, s; u)\{A + \partial_y f(s, y(s, u))\}y$$

(see [8] for details). Accordingly, if we define a control by $v_n(s) = \{y - s(A + \partial_y f(s, y(s, \tilde{u}^n)))y\}/t$, then

$$\int_0^{t_n} S(t_n, s; \tilde{u}^n)Bv_n(s)\, ds = y. \tag{6.35}$$

Since $v(s) \in W(0, t_n; U)$ for sufficiently small $\|Ay\|$, it follows that the sequence of sets $A(K(0, t_n, \tilde{u}^n; U) \cap D(A))$ contains a common open set.

The proof ends applying Lemma 6.3 in the space $H = \mathbb{R} \times D(A)$, where $D(A)$ is endowed with the graph norm. The statement just proved above is that the sequence of sets $K(0, t_n, \tilde{u}^n; U) \cap D(A)$ contains a common open set in the space H. On the other hand, the boundedness assumption for elements of the form (6.34) implies that the sequence $\tilde{K}(0, t_n, \tilde{u}^n; U) \cap (\mathbb{R} \times D(A))$ is bounded in $\mathbb{R} \times D(A)$ (that elements of the form (6.33) are bounded is a consequence of the other hypotheses). Hence, $K(0, t_n, \tilde{u}^n; U) \cap (\mathbb{R} \times D(A))$ is of finite codimension there; in view of the convergence relation proved in Lemma (6.3), the same is true of the sequence $K(0, t_n, \tilde{u}^n; U) \cap (\mathbb{R} \times D(A)) - (0, \Lambda(t_n)\tilde{u}^n)$, thus Theorem 6.4 follows from (6.17) and Lemma 6.3.

A separate statement is needed for the time optimal problem. The function under consideration here is not (6.3) but

$$F_n(u) = \|Ay(t_n, u) - A\bar{y}\| \tag{6.36}$$

in the space $W(0, t_n; U)$. Arguing in the same way as with (6.5) we obtain the following theorem.

THEOREM 6.6 (*The modified sequence maximum principle for the time optimal problem*). *Let $\{u^n\}$ be a sequence of (A, t_n, u^n)-suboptimal controls with $t_n < \bar{t}$, $\varepsilon_n \to 0$. Then there exists a subsequence of $\{u^n\}$ (which we denote by the same symbol) such that*

$$t_n \to \tilde{t} \leqq \bar{t}, \qquad Ay(t_n, u^n) \to A\bar{y}, \tag{6.37}$$

a second sequence $\{\tilde{u}^n\}$ of controls with $\tilde{u}^n \in W(-k, t_n; U)$ such that

$$F_n(\tilde{u}^n) < F_n(u^n) < \infty, \tag{6.38}$$

$$d_n(u^n, \tilde{u}^n) \to 0, \tag{6.39}$$

a sequence $\{y^n\} \in D(A)$ *such that* $\|Ay^n\| = 1$ *and a set* e_n *of full measure in* $0 \leqq s \leqq t_n$ *such that*

$$\langle Ay^n, AS(t_n, s; \tilde{u}^n)B(v - \tilde{u}^n(s))\rangle \geqq -\delta_n \to 0 \quad (v \in U, s \in e_n). \tag{6.40}$$

The following analogue of Theorem 6.4 is proved essentially in the same way.

THEOREM 6.7. *Let the semigroup* $S(t)$ *generated by* A *be analytic, and let the nonlinearity be of the form* (5.1), *with* $B = I$ *and* $f(t, y)$ *satisfying assumptions* (F_α). *Let the control set* U *be closed and bounded and have* 0 *as an interior point. Then the sequence* $\{y^n\}$ *in Theorem* 6.5 *has a subsequence weakly convergent in* $D(A)$ *to* $y \neq 0$.

It can be proved (much in the same way as for the convergence principles in § 5) that the generalized convergence principles just established imply strong convergence of sequences of suboptimal controls. Again we consider only the time optimal case.

We will need the following two results, where the assumptions on A and on $f(t, y)$ in Theorem 6.7 are in force.

LEMMA 6.8. *Let* $h > 0$. *The* (E)*-valued function*

$$(t, s, u) \to AS(t, s; u) \tag{6.41}$$

is uniformly continuous (*in the uniform topology of operators*) *in* $0 \leqq s \leqq t - h \leqq T$, $u \in W(0, T; U)$ ($W(0, T; U)$ *endowed with the weak* $L^2(0, T; E)$*-topology.*

LEMMA 6.9. *Let* $0 \leqq s < t \leqq T$, $u \in W(0, t; U)$. *Then the operator* $AS(t, s; u) : E \to E$ *is compact.*

The proofs of both results are consequences of Lemma 5.4 and Lemma 5.5. We use the integral equation (4.11) premultiplied by A:

$$\begin{aligned} AS(t, s; u)z &= AS(t - s)z \\ &+ \int_s^t (-A)^{1-\alpha} S(t - \sigma)(-A)^\alpha \partial_y f(\sigma, y(\sigma, u))S(\sigma, s; u)z \, ds. \end{aligned} \tag{6.42}$$

Lemma 6.8 follows immediately from the estimate (6.2), from the assumptions on $f(t, y)$, from the fact that $AS(t)$ is continuous in the uniform topology of operators in $t > 0$ and from Lemma 5.4, which guarantees that $S(\sigma, s; u)$ is continuous in the uniform topology of operators in $s < \sigma < T$; we divide the domain of integration in (6.42) in three parts, one $s + \delta < t < T - \delta$ where (E)-continuity of the integrand can be exploited, and two residual intervals, $s < s + \delta$ and $T - \delta < T$, which yield an integral as small as desired. To prove Lemma 6.9 we argue in the same way, using the fact that $AS(t) = S'(t)$ is compact for each $t > 0$ and that $S(\sigma, s; u)$ is compact in $s < \sigma < T$ (Lemma 5.5).

We rewrite (6.40) in the form

$$\langle (AS(t_n, s; \tilde{u}^n))^* Ay^n, v - \tilde{u}^n(s)\rangle \geqq -\delta_n \to 0 \quad (v \in U, s \in e_n). \tag{6.43}$$

Consider the following analogue of (5.12):

$$(AS(\bar{t}, s; \tilde{u}^n))^* Ay^n = ((AS(t_n, s; \tilde{u}^n))^* - (AS(\bar{t}, s; \bar{u})))^* y^n + (AS(\bar{t}, s; \bar{u}))^* y^n. \tag{6.44}$$

Using Lemmas 6.7 and 6.8, we deduce that

$$z_n(t) = (AS(t_n, s; \tilde{u}^n))^* Ay^n \to z(t) = (AS(\bar{t}, s; \bar{u}))^* Ay \quad \text{strongly} \tag{6.45}$$

(where $y \neq 0$ is the weak limit of y^n in the space $D(A)$), so that

$$\langle z(t), v - \tilde{u}^n(s)\rangle \geqq -\delta_n \to 0 \quad (v \in U, s \in e_n). \tag{6.46}$$

From then on, all the arguments following (5.14) apply; the only difference is that $z(t)$ was continuous in $0 \leqq t \leqq \bar{t}$ there, while it is only continuous in $0 \leqq t < \bar{t}$ here, but this does not introduce any significant changes.

7. The strong convergence principle: the noncompact case. The weak convergence principle, as defined in § 4, is based on two sets of hypotheses that work in opposite directions. On the one hand, compactness (either of $S(t)$ or of B) is needed; on the other hand, results such as Lemma 6.3 require, roughly speaking, *exact controllability* (of the linearized system), which is unattainable (except in finite-dimensional situations) when $S(t)$ or B are compact (for the latter case, see [6]). Thus, only modified versions of the convergence principle work in very particular situations, such as the one treated in § 6 under the assumption of compactness of $S(t)$. Compactness of B works in the case (not treated in this paper) of hyperbolic equations, where exact controllability to certain (energy) subspaces holds; however, the requirements on B essentially reduce us to the case of one space variable.

A way out may be attempted giving up the compactness properties of $B^*S(t, s; u)^*$ altogether and trying to establish instead a strong convergence principle. However, this only works in very particular situations, as seen in this section and in § 8; on the other hand, compactness in a weakened form is still necessary.

We consider again systems defined by the initial value problem (4.3)-(4.4) in a Hilbert space E. We do not assume that A is compact, but only that $R(\mu; A) = (\mu I - A)^{-1}$ is compact for some μ in the resolvent set $\rho(A)$. The nonlinear term $f(t, y, u)$ is assumed to be of the form (5.1) and satisfy assumptions (F) in § 4.

LEMMA 7.1. *Let $S(t)$ be a strongly continuous semigroup such that $R(\mu; A)$ is compact for some $\mu \in \rho(A)$. Let the operator $\Pi: L^2(0, T; E) \to C(0, T; E)$ be defined by*

$$(\Pi u)(t) = \int_0^t S(t-\sigma)u(\sigma)\, d\sigma. \tag{7.1}$$

Let $\{u^n\}$ be a sequence in $L^2(0, T; E)$ converging weakly to an element $u \in L^2(0, T; E)$. Then (a)

$$\Pi u^n \to \Pi u \quad \textit{strongly in } L^2(0, T; E). \tag{7.2}$$

(b) *If $\{t_n\}$ is a sequence in $[0, T]$ such that $t_n \to t_0$, then*

$$(\Pi u_n)(t_n) \to (\Pi u)(t_0) \quad \textit{weakly in } E. \tag{7.3}$$

A sketch of the proof is in [10, Lemma 5.1]. We can write Π as the convolution $\Pi u = S * u$, where we have set $S(t) = 0$ in $t < 0$. Translating A if necessary, we may assume as well that (6.1) holds for some $\beta > 0$. Denoting by Φ the Fourier transform operator, we have $(\Phi\Pi u)(\sigma) = R(-i\sigma; A)\Phi u(\sigma)$. For $u \in E$ we have

$$\langle \Phi u_n(\sigma), u\rangle = \int_0^T \langle u_n(t), e^{-i\sigma t}u\rangle\, dt \to \int_0^T \langle u(t), e^{-i\sigma t}u\rangle\, dt = \langle \Phi u(\sigma), u\rangle,$$

(where we have set $u(t) = u_n(t)$ in $t > T$) so that $\Phi u_n(\sigma) \to \Phi u(\sigma)$ weakly in E for each σ. Since $R(\mu; A)$ is compact for some μ, it follows from the second resolvent equation that $R(-i\sigma; A)$ is compact for every σ, so that $R(-i\sigma; A)\Phi u_n(\sigma) \to R(-i\sigma; A)\Phi u(\sigma)$ strongly in E for $-\infty < \sigma < \infty$. Noting that by (6.1), $R(-i\sigma; A)$ is uniformly bounded, (a) follows from the dominated convergence theorem and Plancherel's theorem. To prove (b) we only have to observe that, for $u \in E$,

$$\begin{aligned} \langle (\Pi u_n)(t_n), u\rangle &= \int_0^{t_n} \langle u_n(t), S(t_n - t)u\rangle\, dt \to \int_0^{t_n} \langle u(t), S(t-t)u\rangle\, dt \\ &= \langle (\Pi u)(t), u\rangle, \end{aligned} \tag{7.4}$$

noting that $S(t_n - t)u \to S(t-t)u$ strongly in $L^2(0, T; E)$. Of course, in this last step, compactness of the resolvent $R(\mu; A)$ is unnecessary.

The following result is the analogue of Theorem 5.2 for the present hypotheses. As in § 5 we assume that the control set U is closed and bounded, and we denote by $V = \overline{\text{conv}}\,(U)$ the closed convex hull of U.

THEOREM 7.2. *Let the semigroup $S(t)$ generated by A satisfy the assumptions of Lemma 7.1, and suppose $f(t, y, u)$ is of the form (5.1) with $f(t, y)$ satisfying (F). Let $\{u^n\}$ be a sequence of controls with $u^n \in W(0, t_n; U)$. Assume that $t_n \to t_0$. Then there exists a subsequence of $\{u^n\}$ (which we denote by the same symbol) and a $\bar{u} \in W(0, t_0; V)$ such that*

$$u^n(t) \to \bar{u}(t) \quad \textit{weakly in } L^2(0, t_0; F), \tag{7.5}$$

$$y(t, u^n) \to y(t, \bar{u}) \quad \textit{strongly in } L^2(0, T; U), \tag{7.6}$$

$$y(t_n, u^n) \to y(t_0, \bar{u}) \quad \textit{weakly in } E. \tag{7.7}$$

Again, a sketch of the proof can be found in [10, Thm. 5.2]. We use the successive approximation equation (5.8) to compute $y(t, u^n)$ as the limit of the sequence

$$\{y^m(t, u^n);\ m = 1, 2, \cdots\}. \tag{7.8}$$

We have already observed in § 5 that $\{y^m(t, u^n)\}$ converges absolutely and uniformly (with respect to σ and n) as $m \to \infty$ in some interval $[0, T_0]$ independent of n. The (analogously defined) sequence

$$\{y^m(t, \bar{u});\ m = 1, 2, \cdots\} \tag{7.9}$$

used to compute $y(t, \bar{u})$ enjoys the same convergence properties.

Select a subsequence of $\{u^n\}$ (denoted with the same symbol) such that $u^n \to \bar{u} \in L^2(0, T; F)$ weakly; since $W(0, T; V)$ is convex and closed (hence weakly closed) in $L^2(0, T; F)$, $\bar{u} \in W(0, T; V)$. Making use of (5.8) and of Lemma 5.1 and passing to a subsequence, we deduce that $\{y^1(t, u^n)\}$ is strongly convergent in $L^2(0, T_0; F)$ and convergent almost everywhere in $[0, T_0]$. Using then the dominated convergence theorem in the first term, Lemma 5.1 in the second term and passing to a subsequence, we deduce that $y^1(t, u^n)$ converges strongly in $L^2(0, T_0; F)$ and almost everywhere in $[0, T_0]$ to $y^1(t, \bar{u})$. Operating inductively in the same fashion (using at each step the dominated convergence theorem in the first term of (5.8) and Lemma 7.1 in the second term, and taking a diagonal subsequence at the end) we show that each of the terms in the sequence $\{y^m(t, u^n)\}$ converges in $L^2(0, T_0; F)$ to the corresponding term of the sequence $\{y^m(t, \bar{u})\}$. Noting that

$$\begin{aligned} y(t, u^n) - y(t, \bar{u}) = {} & (y(t, u^n) - y^m(t, u^n)) + (y^m(t, u^n) - y^m(t, \bar{u})) \\ & + (y^m(t, \bar{u}) - y(t, \bar{u})) \end{aligned} \tag{7.10}$$

and using the convergence properties just mentioned, (7.6) follows in the interval $[0, T_0]$. To extend the result to all the interval $[0, t_0]$ we proceed as follows. Since $y(t, u^n) \to y(t, \bar{u})$ almost everywhere in $[0, T_0]$, we may assume (if needed shifting T_0 to the left) that $y(T_0, u^n) \to y(T_0, \bar{u})$. To solve in $t \geqq T_0$, we use the approximation scheme (4.6), that is

$$\begin{aligned} y^{m+1}(t, u^n) = {} & S(t)y(T_0, u^n) \\ & + \int_{T_0}^{t} S(t-\sigma) f(\sigma, y^m(\sigma, u^n))\, d\sigma + \int_{T_0}^{t} S(t-\sigma) B u^n(\sigma)\, d\sigma \end{aligned} \tag{7.11}$$

in $t \geqq T_0$. Arguing in the same way (and, of course, passing to a subsequence), we show that $y(t, u^n)$ converges in $L^2([T_0, T_1]; F)$ and almost everywhere in $[T_0, T_1]$ to

$y(t; \bar{u})$. Using the same argument in intervals $[T_1, T_2], [T_2, T_3], \cdots$ whose length does not tend to zero because of (4.8) (see the arguments following (4.8)), we obtain (7.6) in the whole interval $[0, t_0]$.

The proof of (7.7) is similar. We start with the first equation (7.14) ($m = 0$) and use Lemma 7.1(b) in order to show that $y^1(T_0, u^n) \to y^1(T_0, \bar{u})$ weakly in E. We then make use of (7.11) inductively in the interval $[0, T_0]$, combined with the fact that, passing to a subsequence we may assume that each sequence $\{y^m(t, u^n)\}$ converges almost everywhere in $[0, T_0]$ as $n \to \infty$ to $y(t, \bar{u})$ to show that, for each $m = 1, 2, \cdots$

$$y^m(T_0, u^n) \to y^m(T_0, \bar{u}) \quad \text{weakly in } E.$$

This, combined with the convergence properties as $m \to \infty$ of the sequence (7.8), shows that

$$y(T_0, u^n) \to y(T_0, \bar{u}) \quad \text{weakly in } E. \tag{7.12}$$

Using then (7.11) and arguing in the same way, we show (7.7) in the intervals $[T_1, T_2]$, $[T_2, T_3], \cdots$ until the limit t_0 and the sequence $\{t_n\}$ are eventually contained in an interval $[T_k, T_{k+1}]$.

Remark 7.3. Using an argument similar to that in Remark 5.3 we can show that it is unnecessary to take a subsequence in Theorem 7.2.

LEMMA 7.4. *Under the assumptions on $S(t)$ and $f(t, y)$ in Theorem* 7.2 *the (E)-valued functions*

$$(t, s, u) \to S(t, s; u) \tag{7.13}$$

$$(t, s, u) \to S(t, s; u)^* \tag{7.14}$$

are uniformly continuous (in the strong topology of operators) in $0 \leqq s \leqq t \leqq T$, $u \in W(0, T; U)$ ($W(0, T; U)$ endowed with the weak $L^2(0, T; F)$-topology).

The proof (which is also sketched in [10, Thm. 5.3]) uses the integral equation (4.11) defining $S(t, s; u)$, the integral equation

$$S(t, s; u)^* z = S(t-s)^* z + \int_s^t S(\sigma, s; u)^* B(\sigma)^* S(t-\sigma)^* z \, d\sigma \tag{7.15}$$

$$(0 \leqq s \leqq t \leqq T)$$

obtained from (4.11) taking adjoints, the L^2-convergence of $\{y(t, u^n)\}$ and the uniform boundedness of $\|y(t, u^n)\|$. We omit the details.

The following result on the operator $\Lambda(t)u$ defined by (6.26) roughly corresponds to Lemma 6.4.

LEMMA 7.5. *Let $\{\tilde{u}^n\}$ be the sequence in* (5.3). *Then there exists a subsequence (denoted with the same symbol) such that $\{\Lambda(t_n)\tilde{u}^n\}$ is strongly convergent in E.*

The proof is based on (6.30) for $t = t_n$, $u = \tilde{u}^n$. Obviously, the left-hand side and the first term on the right-hand side converge, the latter because of Lemma 7.4. In the integrand we again use Lemma 7.4 for $S(t, \sigma; u)$, uniform boundedness of $\{y(t, u^n)\}$, pointwise convergence of the same sequence (which can be achieved passing to a subsequence) and the dominated convergence theorem.

At this point, there is no difficulty in upgrading the sequence maximum principles in § 5 into weak convergence principles, that is, in establishing the weak convergence of (a subsequence of) the sequence $\{y^n\}$ to a limit $y \neq 0$. However, this would not be useful in the present situation, since the crucial uniform continuity-compactness properties of the operator $S(t, s; u)$, present in the abstract parabolic case, are now lacking. Thus, we need to establish strong convergence of a subsequence of $\{y^n\}$, which we do in the next section.

8. The convergence principle: the noncompact case. Let Δ be an arbitrary set in a Hilbert space H, and let $S(y, \rho)$ be the closed sphere of center y and radius ρ. We say that a point y in the space H is Δ-*regular* if and only if there exists a hyperplane

$$H(y, \bar{y}) = \{v; \langle v - y, \bar{y}\rangle = 0\} \tag{8.1}$$

(where we assume that the *generator* $\bar{y}$ of $H(y, \bar{y})$ satisfies $\|\bar{y}\| = 1$), a number $\rho > 0$ and a map

$$\Pi : S(y, \rho) \cap H(y, \bar{y}) \to \Delta \tag{8.2}$$

such that

$$\|\Pi(z) - z\| \leqq r(\|z - y\|) \quad (\|z - y\| \leqq \rho), \tag{8.3}$$

where the function $r(\mu)$ is defined in $0 < \mu \leqq \rho$ and satisfies

$$r(\mu) = 0(\mu) \quad \text{as } \mu \to 0. \tag{8.4}$$

Example 8.1. Let $\Delta = B(0, R)$ be the sphere of center 0 and radius R in H, and let y be a point in H. If $\|y\| > R$ then it is obvious that y is not Δ-regular. If $\|y\| < R$ then y is Δ-regular (we may take $\rho = R - \|y\|$ and $H(y, \bar{y})$ an arbitrary hyperplane). The only nontrivial situation is that where y belongs to the boundary of Δ, that is, $\|y\| = R$. In this case we take $\bar{u} = y/\|y\|$ and define the map Π by

$$\Pi(z) = (R/\|z\|)z. \tag{8.5}$$

Let $z \in H(y, \bar{y})$. Trigonometry in the two-dimensional subspace generated by y and z reveals that $\|z - y\|/R = \tan\theta$, $\|z - y\|/\|z\| = \sin\theta$, where θ is the angle between y and z. Accordingly,

$$\|\Pi(z) - z\| = \|(R/\|z\|)z - z\| = \|z\| - R = \{(1 - \cos\theta)/\sin\theta\}\|z - y\|, \tag{8.6}$$

and it follows that y is Δ-regular.

LEMMA 8.2. *Let y be a Δ-regular point, and let $\{y^n\}$ be a sequence in H such that* $\|y^n\| = 1$,

$$\langle y^n, z - y\rangle \geqq -\varepsilon_n \to 0 \quad (z \in \Delta). \tag{8.7}$$

Then

$$y^n \to \bar{y} \text{ strongly in } H, \tag{8.8}$$

where $\bar{y}$ is the generator of the hyperplane $H(y, \bar{y})$ in (8.1).

Proof. Let $z \in H(y, \bar{y})$. We have

$$\begin{aligned}\langle y^n, z - y\rangle &= \langle y^n, \Pi(z) - y\rangle - \langle y^n, \Pi(z) - z\rangle \geqq -\varepsilon_n - \|\Pi(z) - z\| \\ &\geqq -\varepsilon_n - r(\|z - y\|).\end{aligned}$$

Let $\delta > 0$. Choose μ so small that $r(\mu)/\mu < \delta/2$ and then n so large that $\varepsilon_m/\mu < \delta/2$ for $m \geqq n$. Taking $\|z - y\| = \mu$ and writing $w = (z - y)/\|z - y\|$ we obtain the relation

$$\langle y^m, w\rangle \geqq -\delta \quad (\|w\| = 1, \langle w, \bar{y}\rangle = 0, m \geqq n).$$

This is easily seen to imply (8.8).

In the following result we consider a sequence $\{\Delta_n\}$ of sets in H and a sequence $\{y_n\}$ in H. We say that $\{y_n\}$ is $\{\Delta_n\}$-*regular* if and only if there exists a sequence $\{H(y_n, \bar{y}_n)\}$ of hyperplanes (with generators $\bar{y}_n$, $\|\bar{y}_n\| = 1$) and a sequence of maps

$$\Pi_n : S(y_n, \rho) \cap H(y_n, \bar{y}_n) \to \Delta_n \tag{8.9}$$

such that

$$\|\Pi_n(z_n) - z_n\| \leqq r(\|z_n - y_n\|)(\|z_n - y_n\| \leqq \rho), \tag{8.10}$$

with

$$r(\mu) = o(\mu) \quad \text{as } \mu \to 0, \tag{8.11}$$

where $r(\mu)$ is defined in $0 < \mu < \rho$. The definition of course implies (but is not equivalent to) the fact that each y_n is Δ_n-regular; note that ρ and $r(\mu)$ are assumed to be independent of n.

LEMMA 8.3. *Let $\{\Delta_n\}$ be a sequence of sets in the Hilbert space H, and let $\{y_n\}$ be a $\{\Delta_n\}$-regular sequence. Assume the sequence $\{\bar{y}_n\}$ of generators of the cones $H(y_n, \bar{y}_n)$ satisfies*

$$\bar{y}_n \to \bar{y} \quad \textit{strongly in } H. \tag{8.12}$$

Let $\{y^n\}$ be a sequence such that $\|y^n\| = 1$ and

$$\langle y^n, z_n - y_n\rangle \geqq -\varepsilon_n \to 0 \quad (z_n \in \Delta_n). \tag{8.13}$$

Then

$$y^n \to \bar{y} \quad \textit{strongly in } H. \tag{8.14}$$

The proof is an obvious generalization of the proof of Lemma 8.2. Let $z_n \in H(y_n, \bar{y}_n)$. Then we have

$$\begin{aligned}\langle y^n, z_n - y_n\rangle &= \langle y^n, \Pi_n(z_n) - y_n\rangle - \langle y^n, \Pi_n(z_n) - z_n\rangle \geqq -\varepsilon_n - \|\Pi_n(z_n) - z_n\| \\ &\geqq -\varepsilon_n - r(\|z_n - y_n\|).\end{aligned}$$

We again take $\delta > 0$ and choose μ so small that $r(\mu)/\mu < \delta/2$ and then n so large that $\varepsilon_m/\mu < \delta/2$ for $m \geqq n$. Taking $\|z_m - y_m\| = \mu$ and writing $w_m = (z_m - y_m)/\|z_m - y_m\|$ we obtain the relation

$$\langle y^m, w_m\rangle \geqq -\delta \,(\|w_m\| = 1, \langle w_m, \bar{y}_m\rangle = 0, m \geqq n), \tag{8.15}$$

which is easily seen to imply (8.14).

In what follows, we work under the assumptions in § 8, that is, we require $R(\mu; A)$ to be compact for some $\mu \in \rho(A)$. We prove a convergence principle pertaining to the time optimal problem.

THEOREM 8.4. *Assume that $R(\mu; A)$ is compact for some $\mu \in \rho(A)$ and that $f(t, y, u)$ is of the form* (5.1) *with $f(t, y)$ satisfying conditions (F) in § 4. Assume that, for every convergent sequence $\{t_n\}$ and every weakly convergent sequence $\{\tilde{u}^n\}$ in $W(0, t_n; U)$ the sequence $\{\Lambda\tilde{u}^n(t_n)\}$ is $\{K(0, t_n, \tilde{u}^n; U)\}$-regular. Assume further that*

$$\bar{y}_n \to \bar{y}, \tag{8.16}$$

where $\bar{y}_n$ is the generator of the hyperplane $H(y_n, \bar{y}_n)$ in (8.9). *Then the sequence $\{y^n\}$ in the sequence maximum principle* (5.3) *is strongly convergent.*

The proof is an immediate consequence of Lemma 8.3.

The $\{K(0, t_n, \tilde{u}^n; U)\}$-regularity assumption in Theorem 8.4 is difficult to verify, since very little is known about reachable sets $\{K(0, t_n, \tilde{u}^n; U)\}$ in infinite-dimensional spaces. An example where $\{K(0, t_n, \tilde{u}^n; U)\}$-regularity holds is the following.

COROLLARY 8.5. *Let A and f satisfy the conditions in Theorem* 8.5. *Assume that*

$$B = I, \quad U = B(0, R) = \{u \in E; \|u\| \leqq R\}. \tag{8.17}$$

Assume in addition that $A + \partial_y f(t, y(t, u))$ is skew adjoint for every t in $0 \leqq t \leqq T$ and every $u \in W(0, T; U)$. Then the sequence $\{y^n\}$ in the sequence maximum principle (5.3) *is strongly convergent.*

Proof. Since each $A+\partial_y f(t, y(t, u))$ is skew adjoint, the solution operator $S(t, s; u)$ is unitary (isometric and invertible) for each s, t. (In particular, $\|S(t, s; u)\| = 1$.) It follows that, for each $\bar{t}$ and each $u \in W(0, \bar{t}; U)$ we have

$$K(0, \bar{t}, u; U) = \bar{t}B(0, R) = B(0, \bar{t}R). \tag{8.18}$$

In fact, it is obvious that every element of $K(0, \bar{t}, u; U)$ has norm $\leqq \bar{t}R$; on the other hand, if $v \in B(0, \bar{t}R)$ then $K(0, \bar{t}, u; U)$, since

$$v = \int_0^{\bar{t}} S(t, s; u)u(s)\, ds$$

with $u(s) = \bar{t}^{-1}S(t, s; u)^{-1}v = \bar{t}^{-1}S(s, t; u)v$.

We show that the sequence $\{\Lambda\tilde{u}^n(t_n)\}$ is $\{K(0, t_n, \tilde{u}^n; U)\}$-regular. We know from Lemma 7.6 that $\{\Lambda\tilde{u}^n(t_n)\} \to y$ strongly; under the present hypotheses $\|\Lambda\tilde{u}^n(t_n)\| \leqq \bar{t}R$, so that $\|y\| \leqq \bar{t}R$. We argue essentially as in Example 8.1. If $\|y\| < \bar{t}R$ then we may use the sequence of hyperplanes $H(y_n, \bar{y})$ with $\bar{y}$ arbitrary and Π the identity map. When $\|y\| = R$ we take $\bar{y}_n = y_n/\|y_n\|$ and we define $\Pi(z) = Rz/\|z\|$. Using (8.6) for $\theta = \theta_n$ the angle between y and z, (8.10) follows. This ends the proof.

By way of conclusion, we note that the difficulties in proving convergence principles (and, through them, strong convergence of optimal controls in the point target case) are essentially the same as those difficulties involved in proving the maximum principle with point target (see [8]). The results available make evident that in infinite-dimensional spaces, either the maximum principle or convergence of suboptimal controls will only hold in very particular situations, two of which were treated in this paper. A way out of this difficulty is to give slackness to the target condition (as done in [8] for the maximum principle and in [9], [10] for convergence of suboptimal controls).

For a different approach to the convergence problem see [12], which covers some cases where the maximum principle does not hold ([5]).

Acknowledgments. The author is grateful to H. Frankowska for many improvements in the presentation, in particular for pointing out the correct function $F_n(u)$ in (3.4). Also, the referee's observations resulted in a much improved paper.

REFERENCES

[1] V. BARBU, *Optimal Control of Variational Inequalities*, Research Notes in Mathematics 100, Pitman, London, 1984.

[2] ———, *The time optimal problem for a class of nonlinear distributed systems*, in Control Problems for Systems Described by Partial Differential Equations and Applications, Springer Lecture Notes in Control and Information Sciences 97, 1987, pp. 16-39.

[3] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[4] I. EKELAND, *Nonconvex minimization problems*, Bull. Amer. Math. Soc., 1 (1979), pp. 443-474.

[5] H. O. FATTORINI, *The time optimal control problem in Banach spaces*, Appl. Math. Optim., 1 (1974), pp. 163-188.

[6] ———, *Exact controllability of linear systems in infinite dimensional spaces*, in Partial Differential Equations and Related Topics, Springer Lecture Notes in Mathematics 446, 1975, pp. 166-183.

[7] ———, *The maximum principle for nonlinear nonconvex systems in infinite dimensional spaces*, in Distributed Parameter Systems, Springer Lecture Notes in Control and Information Sciences 75, 1986, pp. 162-178.

[8] ———, *A unified theory of necessary conditions for nonlinear nonconvex control systems*, Appl. Math. Optim., 15 (1987), pp. 141-185.

[9] ———, *Optimal control of nonlinear systems: convergence of suboptimal controls*, I, in Operator Methods for Optimal Control Problems, Lecture Notes in Mathematics 108, Marcel Dekker, New York, Basel, 1987, pp. 159-199.

[10] H. O. FATTORINI, *Optimal control of nonlinear systems: convergence of suboptimal controls*, II, in Control Problems for Systems Described by Partial Differential Equations and Applications, Springer Lecture Notes in Control and Information Sciences 97, 1987, pp. 230–246.

[11] ———, *Convergence of suboptimal controls for point targets*, in Optimal Control of Partial Differential Equations II, International Series on Numerical Mathematics 78, Birkhäuser Verlag, Basel, 1987, pp. 91–107.

[12] ———, *Some Remarks on Convergence of Suboptimal Controls*, A. V. Balakrishnan Anniversary Volume, Software Optimization, New York, 1987, pp. 144–148.

[13] V. P. MIHAILOV, *Partial Differential Equations*, Mir, Moscow, 1978.

[14] V. I. PLOTNIKOV AND M. I. SUMIN, *The construction of minimizing sequences in problems of control of systems with distributed parameters*, Zh. Vychisl. Mat. i Mat. Fiz., 22 (1982), pp. 49–56.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 342-358, March 1990

005

WELL-POSEDNESS OF H^∞ OPTIMAL CONTROL PROBLEMS*

MALCOLM C. SMITH†

Abstract. This paper considers the effect of perturbations of a nominal single-input/single-output linear plant, or a band of uncertainty, on the solution to certain frequency domain optimal control problems. Weighted H^∞ sensitivity minimization, mixed, and robust sensitivity minimization will be considered along with problems of more general type. A brief discussion of H^p optimal control problems will also be given. Typical examples will be presented where optimal sensitivity does not depend continuously on the plant (ill-posedness) and conditions for well-posedness will be given. It is demonstrated that similar discontinuities can occur for more general problems. Also suggested are ways of defining an optimization problem so that continuous dependence of the infimum is ensured.

Key words. feedback, frequency domain design, sensitivity minimization, H^∞-optimization, uncertainty, perturbations, continuity, approximation, delay systems

AMS(MOS) subject classifications. 93B50, 93B35, 93B28

NOTATION

$D(\bar{D})$	open (closed) unit disc
∂D	unit circle
$\mathbf{C}_+(\bar{\mathbf{C}}_+)$	open (closed) right half plane
$\tilde{\mathbf{C}}_+$	$\bar{\mathbf{C}}_+ \cup \{\infty\}$
$H(X)$	the holomorphic functions on X
$H^p(X)$	the standard Hardy p-space ($1 \leqq p \leqq \infty$) on X (if not explicitly stated H^p means $H^p(\mathbf{C}_+)$)
L^p	the standard Lebesgue p-spaces
$\tilde{H}^p(X)$	$\{f \in H^p(X): \overline{f(s)} = f(\bar{s})\}$
$R\tilde{H}^\infty$	the rational functions in $\tilde{H}^\infty$
$\tilde{A}^\infty$	$\{f \in \tilde{H}^\infty(\mathbf{C}_+)$: f is continuous on $\bar{\mathbf{C}}_+\}$
$\tilde{H}^\infty_-$	$\{f \in \tilde{H}^\infty(X_\sigma)$ for some $\sigma > 0$, where X_σ is the half plane $\mathrm{Re}\,(s) \geqq -\sigma\}$
$\mathbb{R}_{pr}(s)$	the proper (i.e., no poles at ∞) rational functions
$\mathbb{R}_{sp}(s)$	the strictly proper (i.e., having a zero at ∞) rational functions

1. Introduction. The H^∞ framework was introduced into control by Zames [16] principally because it is natural for studying problems involving uncertainty. In classical theory, plant uncertainty is represented by a tolerance band on the frequency response, which is exactly a "weighted" ball in H^∞. For the problem of optimal disturbance attenuation there is a natural H^∞ formulation for disturbances whose spectra are not fixed [17], and many other problems involving sensitivity and robustness optimization can be posed in an H^∞ framework [5]. To date, much work on H^∞ optimization has been devoted to obtaining explicit solutions to various classes of control problems. Contributions have been made to understanding the capabilities and limitations of feedback, and new possibilities have opened up for design.

An aspect of H^∞ optimization which has remained unexplored to date is the question of *well-posedness*. By well-posedness we mean roughly that a vanishingly small change in the problem specification will result in a vanishingly small change in the solution. Of course, this has to be defined more precisely in particular cases, but we will mostly be concerned with the effect of perturbations of the nominal plant (or the tolerance band) on the performance measure and the optimal control. After an initial

* Received by the editors July 15, 1987; accepted for publication (in revised form) April 24, 1989.

† Department of Electrical Engineering, Ohio State University, Columbus, Ohio 43210. This work was supported in part by the Science and Engineering Research Council of the United Kingdom and by the Natural Sciences and Engineering Research Council of Canada.

study it seems that the H^∞ formulation is rather susceptible to the phenomena of ill-posedness. It is the purpose of this paper to point out and analyse some of the circumstances when this occurs and to suggest remedies. This paper is an expanded version of [12].

Before proceeding it is worth giving reasons why we consider well-posedness to be an important property. Clearly, well-posedness is not the same as the ordinary notion of robustness, which requires the computed controller to perform satisfactorily for some neighbourhood of the nominal plant. The latter property must be satisfied as a matter of course. First, we can see that well-posedness is desirable purely from a computational point of view. Obviously if the problem solution is critically dependent on small changes in the plant parameters it will be very difficult to obtain satisfactory computational procedures. Second, it is clear that ill-posedness could be caused by neglecting certain constraints in the problem formulation. In this respect a knowledge of how to achieve well-posedness would be a useful guide when selecting a performance criterion. A third motivation is the desire to be able to tune a controller on-line in cases of parameter drift. Assuming that a family of controllers is designed off-line to satisfy certain performance criteria, a successful implementation will undoubtedly require that the mapping from plant to controller defined by the design problem be continuous.

The paper is organized as follows. In § 2 we consider the problem of H^∞ weighted sensitivity minimization for stable plants. In the finite-dimensional case and for delay systems, typical examples are presented showing discontinuity of the infimum with respect to plant perturbations. A general condition is then derived characterizing plants at which we have continuity (Theorem 1). One important requirement is that the optimum for the plant equals the optimum for the inner (all-pass) part of the plant. We give a general condition for this to hold for a wide class of infinite-dimensional plants (Theorem 2). Finally we show continuity of the optimum with respect to perturbations of the weighting function. In § 3 we obtain similar results for the case of H^p weighted sensitivity minimization. In § 4 we present some ways in which the sensitivity minimization problem can be modified to ensure continuity of the infimum (with respect to plant perturbations) for arbitrary plants in H^∞. In § 5 we consider the problem of minimizing the supremal weighted sensitivity over a ball of plants. We show that the optimum does not depend continuously on the radius of the ball, and suggest a modified problem which has continuous dependence. In § 6 we consider a general design problem for (possibly) unstable plants in which the infimum of a weighted combination of closed loop transfer functions is sought. A sufficient condition is obtained for the optimum to depend continuously on the plant (corollary to Theorem 7).

2. H^∞ weighted sensitivity minimization. We begin by recalling the H^∞ optimal sensitivity problem. Consider a plant $P(s) \in \tilde{H}^\infty + \mathbb{R}_{pr}(s)$ in the standard feedback configuration of Fig. 1 where the controller $F(s) \in \tilde{H}^\infty + \mathbb{R}_{pr}(s)$ is to be designed. The feedback loop is defined to be *stable* if all four closed loop transfer functions $x_i \to e_j$ are L^2 bounded-input/bounded-output stable (i.e., elements of $\tilde{H}^\infty$). This is equivalent to

$$P(1+FP)^{-1}, \quad F(1+PF)^{-1} \in \tilde{H}^\infty. \tag{2.1}$$

In this section (and through to § 5) we assume $P(s) \in \tilde{H}^\infty$, in which case (2.1) is equivalent to

$$Q := F(1+PF)^{-1} \in \tilde{H}^\infty. \tag{2.2}$$

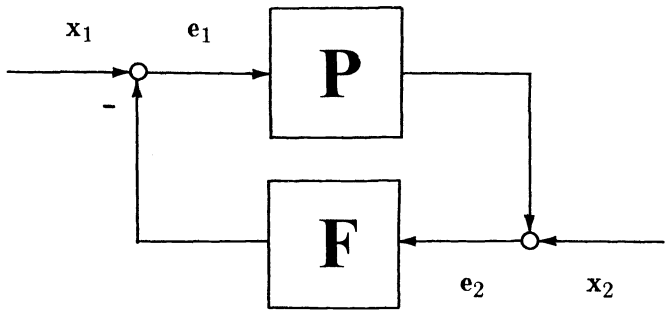


FIG. 1

The H^∞ weighted sensitivity minimization problem is to find

$$\mu(P) := \inf_{F \text{ stablz.}} \| W(1+PF)^{-1} \|_\infty \tag{2.3}$$

where $W(s) \in \tilde{H}^\infty$ is outer and the infimum is taken over all controllers $F(s) \in \tilde{H}^\infty + \mathbb{R}_{pr}(s)$ which stabilize the feedback loop. If $P \in \tilde{H}^\infty$ then (2.3) reduces to

$$\begin{aligned} \mu(P) &= \inf_{\substack{Q \in \tilde{H}^\infty \\ Q/(1-PQ) \in \tilde{H}^\infty + \mathbb{R}_{pr}(s)}} \| W(1-PQ) \|_\infty \\ &\geqq \inf_{Q \in \tilde{H}^\infty} \| W(1-PQ) \|_\infty =: \nu(P). \end{aligned} \tag{2.4}$$

It is usually assumed that (2.4) holds with equality. However, this is not immediate for nonrational P and does not seem to have been established for completely general $P \in H^\infty$. One problem is that $1-PQ$ could have infinitely many $\mathbf{C}_+$ zeros with the consequence that $F = Q/(1-PQ)$ has infinitely many $\mathbf{C}_+$ poles. If W is rational and P is continuous on $\bar{\mathbf{C}}_+$ then equality can be established as in the proof of Theorem 2. For the sake of generality, we will not assume in this paper that (2.4) holds with equality.

We wish to consider the effect of perturbations of P on $\mu(P)$ (and the corresponding optimal F). In particular we want to investigate if the mapping $P \to \mu(P)$ (and $P \to F(P)$) is continuous. Now if $P \in \tilde{H}^\infty$ we have a natural topology on plants defined by the H^∞ norm, which is also physically meaningful (see concluding remarks). In this case our problem reduces to: does $\| P_i - P \|_\infty \to 0$ imply $\mu(P_i) \to \mu(P)$? The answer is clearly no, and the following is a simple example which shows this.

Example 1. Let

$$P(s) = \frac{s}{s+1}, \qquad W(s) = \frac{1}{s+1}.$$

Clearly $\mu(P) \geqq |W(0)| = 1$. In fact $\mu(P) = 1$ and there are a large number of F's which achieve the minimum, e.g., $F = 0$. We note however that $\| P_\varepsilon - P \|_\infty \to 0$ as $\varepsilon \to 0$ where $P_\varepsilon(s) = (s+\varepsilon)/(s+1)$, and that $\mu(P_\varepsilon) = 0$ for all $\varepsilon > 0$. Thus $\mu(\cdot)$ is not continuous at P.

The discontinuity in $\mu(P)$ can be removed in Example 1 by selecting a weighting function $W(s)$ with $W(0) = 0$. However this may be in conflict with a choice which appropriately reflects the spectrum of possible disturbances. There is also no discontinuity if we minimize the 2-norm of the weighted sensitivity rather than the ∞-norm (see § 3). But again there may be strong reasons for preferring a minimax approach. Also, neither of these modifications allows us to choose an optimal F continuously as a function of P. A second instance of discontinuity is the following example.

Example 2. Let

$$P(s)=\frac{e^{-s}}{s+1}, \qquad W(s)=\frac{1}{s+1}.$$

In this case it follows as in [17] (see also Theorem 2) that

$$\mu(P)=\inf_{Q\in\tilde{H}^\infty}\|W-e^{-s}Q\|_\infty.$$

Now from [4] we have $\mu(P)=(1+y_0^2)^{-1/2}$ where y_0 is the unique root of $\tan y+y=0$ lying between $\pi/2$ and π. We note however that $\|P_n-P\|_\infty\to 0$ as $n\to\infty$ where

$$P_n(s)=\frac{1}{(s+1)(1+s/n)^n} \tag{2.5}$$

and clearly $\mu(P_n)=0$ for all n, since P_n is outer (see Theorem 2). Thus $\mu(\cdot)$ is not continuous at P.

Example 2 causes more concern than Example 1 for two reasons. First, there is no obvious way of modifying the weight to ensure even that $\mu(\cdot)$ is continuous. Also the discontinuity in $\mu(\cdot)$ remains when the ∞-norm is replaced by, for example, the 2-norm (see (§ 3). Second, Example 2 represents a plant which is strictly proper and has a delay—this is a very typical situation in practice—whereas a plant which has an imaginary axis zero is more special. Thus Example 2 questions the validity of approximating infinite-dimensional plants by finite-dimensional ones when considering optimal sensitivity (see concluding remarks).

Our first objective is to obtain conditions on P to ensure that $\mu(\cdot)$ is continuous at P. In fact we will show (Theorem 1) that the above examples are the only two types of situations in which $\mu(\cdot)$ can be discontinuous. We begin by establishing the simple fact that $\mu(\cdot)$ is upper semicontinuous on $\tilde{H}^\infty$.

LEMMA 1. *If* P_i, $P\in\tilde{H}^\infty$ *and* $\|P_i-P\|_\infty\to 0$ *as* $i\to\infty$ *then*

$$\limsup \mu(P_i)\leqq\mu(P). \tag{2.6}$$

Proof. Observe that the infimum in $\mu(P)$ can be taken over $Q\in\tilde{H}^\infty$ such that $Q/(1-PQ)\in\tilde{H}^\infty+\mathbb{R}_{pr}(s)$ has no poles on the imaginary axis. For such a Q, $Q/(1-PQ)\in\tilde{H}^\infty+\mathbb{R}_{pr}(s)$ implies that $Q/(1-P_iQ)\in\tilde{H}^\infty+\mathbb{R}_{pr}(s)$ for sufficiently large i. Thus

$$\limsup_i \inf_{\substack{Q\in\tilde{H}^\infty \\ Q/(1-P_iQ)\in\tilde{H}^\infty+\mathbb{R}_{pr}(s)}} \|W(1-P_iQ)\|_\infty \leqq \limsup_i \|W(1-P_i\hat{Q})\|_\infty$$

$$=\|W(1-P\hat{Q})\|_\infty$$

for any $\hat{Q}\in\tilde{H}^\infty$ such that $\hat{Q}/(1-P\hat{Q})\in\tilde{H}^\infty+\mathbb{R}_{pr}(s)$ has no poles on the imaginary axis. The result now follows by taking the infimum over $\hat{Q}$. □

To proceed further we need some standard facts about the factorization of H^∞ functions. We recall that any $P\in\tilde{H}^\infty$ can be written uniquely as $P=BL$ where B, $L\in\tilde{H}^\infty$, B is *inner* or *all-pass* ($|B(j\omega)|=1$ almost everywhere), and L is *outer* or *minimum phase* ($L(s)$ has no zeros in $\mathbf{C}_+$ and $\log|L(s)|$ can be expressed as the Poisson integral of $\log|L(j\omega)|$)—see [6] and [7] for details. Furthermore we can always write $B=B_bB_s$ where B_b, $B_s\in\tilde{H}^\infty$, B_b is a *Blaschke product* and B_s is a *singular inner function.* Transforming the right half plane to the unit disc by setting $z=(s-1)/(s+1)$ we can always write

$$B_b\left(\frac{1+z}{1-z}\right)=\prod_{n=1}^{\infty}\frac{\bar{\alpha}_n}{|\alpha_n|}\left(\frac{\alpha_n-z}{1-\bar{\alpha}_nz}\right)$$

where α_n is a sequence of complex numbers in the disc such that $\Sigma(1-|\alpha_n|)<\infty$. We also have

$$B_s\left(\frac{1+z}{1-z}\right)=\exp\left(-\int\frac{e^{j\theta}+z}{e^{j\theta}-z}\,d\lambda(\theta)\right)$$

where $d\lambda$ is a positive singular measure. We will denote the closed support of the measure by Λ.

Returning to (2.4) we see that

$$\begin{aligned}\nu(P)&=\inf_{Q\in\tilde{H}^\infty}\|W(1-BLQ)\|_\infty\\ &\geqq\inf_{\tilde{Q}\in\tilde{H}^\infty}\|W(1-B\tilde{Q})\|_\infty=\nu(B).\end{aligned}\tag{2.7}$$

Furthermore we can show the following lemma.

LEMMA 2. *For any rational outer* $W(s)\in\tilde{H}^\infty$

$$\nu(B)=\inf_{\hat{Q}\in\tilde{H}^\infty}\|W-B\hat{Q}\|_\infty.\tag{2.8}$$

Proof. This proof generalizes an argument of [17]. Obviously $\nu(B)\geqq\inf\|W-B\hat{Q}\|$. Now if we can find an inverting sequence $L_n\in\tilde{H}^\infty$ such that

$$\|L_nW\|_\infty\leqq 1,\tag{2.9a}$$

$$\|(1-L_nW)W\|_\infty\to 0\quad\text{as } n\to\infty\tag{2.9b}$$

then we can establish (2.8) with equality. To see this observe that for any $\hat{Q}\in\tilde{H}^\infty$

$$W(1-B\hat{Q}L_n)=L_nW(W-B\hat{Q})+(1-L_nW)W$$

and so

$$\|W(1-B\hat{Q}L_n)\|_\infty\leqq\|W-B\hat{Q}\|_\infty+\|(1-L_nW)W\|_\infty,$$

which implies

$$\inf_{\tilde{Q}\in\tilde{H}^\infty}\|W(1-B\tilde{Q})\|_\infty\leqq\inf_{\hat{Q}\in\tilde{H}^\infty}\|W-B\hat{Q}\|_\infty.$$

Now for a rational outer $W(s)$ we can easily find an inverting sequence L_n. Let $W(s)$ have imaginary axis zeros ω_i of multiplicity r_i $(i=1,\cdots,l)$ and r zeros at infinity. Setting

$$L_n(s)=\left(\frac{n}{s+n}\right)^r\prod_{i=1}^{l}\left(\frac{s+j\omega_i}{s+1/n+j\omega_i}\right)^{r_i}W(s)^{-1}$$

we see that $L_n(s)\in\tilde{H}^\infty$ and (2.9a) holds. To verify (2.9b) we note that $L_n(s)W(s)$ tends uniformly to 1 on any compact subset of the imaginary axis not containing any zeros of $W(s)$. □

Remark. This lemma holds for considerably more general outer functions than just rational ones, in fact, for any $W(s)$ for which an inverting sequence satisfying (2.9) exists.

Now (2.8) is a standard mathematical problem which has an extensive literature (see [1], [9], [11]). We can write (2.8) in Nehari problem form as follows:

$$\nu(B)=\inf_{\hat{Q}\in H^\infty}\|WB^*-\hat{Q}\|_\infty=:d(WB^*,H^\infty),$$

where $d(f, H^\infty)$ denotes the distance of the L^∞ function f from H^∞. Now for any $f \in L^\infty$ we can define a multiplication operator $M_f: L^2 \to L^2$ by $M_f x = fx$. If we denote by Π the orthogonal projection from $L^2 \to H^{2\perp}$, the Hankel operator $H_f: H^2 \to H^{2\perp}$ is defined by

$$H_f := \Pi M_{f|H_2}.$$

Now *Nehari's Theorem* [9] says that $d(f, H^\infty) = \|H_f\|$, so $\nu(B) = \|H_{WB^*}\|$. Moreover the infimum in (2.8) is achieved for some $\hat{Q} \in H^\infty$ (not necessarily uniquely).

To date, much work on weighted sensitivity minimization has been concerned with problem (2.8). However it is possible that $\mu(P) > \nu(B)$ and so the optimal sensitivity for P is not obtained by solving (2.8). Example 1 is precisely such a case, since $P(s) = s/(s+1)$ is outer and $\nu(B) = \nu(1) = 0$. Moreover it is clear that the satisfaction of $\mu(P) \geqq \nu(B)$ with equality or inequality is closely tied up with the question of the continuity of $\mu(\cdot)$. In fact, if the outer part L of P is continuous on $\tilde{\mathbf{C}}_+$ then it is easy to see that $\mu(P) > \nu(B)$ implies discontinuity of $\mu(\cdot)$ since L can be uniformly approximated by invertible outer functions. On the other hand it is clear that $\mu(P) = \nu(B)$ is not sufficient for continuity as is shown by Example 2. This example is characterized by the fact that the inner part of the plant is singular *and* $L(s)$ vanishes on the support of the singular measure, namely $s = \infty$. Again $P(s) = e^{-s}/(s+1)$ is continuous on $\tilde{\mathbf{C}}_+$ and can be arbitrarily closely approximated by invertible outer functions. At first sight one might be tempted to conjecture that if e^{-s} is replaced by an infinite Blaschke product B_b then $\mu(\cdot)$ will again be discontinuous, but this is not the case. If $\|P_i - B_b/(s+1)\|_\infty \to 0$ as $i \to \infty$ then the zeros of P_i in $\mathbf{C}_+$ tend to those of B_b by Rouché's theorem. Thus $\liminf \mu(P_i) \geqq \mu(B_{b,T})$ where $B_{b,T}$ is any finite truncation of B_b. But $\sup \mu(B_{b,T}) = \mu(B)$, where the supremum is taken over all truncations (see [15, p. 291, Thm. 9]). Hence $\liminf \mu(P_i) \geqq \mu(B) = \mu(P)$, which together with Lemma 1 establishes continuity.

We are now in a position to derive a general condition for the continuity of $\mu(\cdot)$.

THEOREM 1. *Let $W(s) \in \tilde{H}^\infty$ outer be chosen so that* (2.9) *holds for any B (e.g., $W(s)$ rational). Then $\mu(\cdot)$ is continuous at $P \in \tilde{H}^\infty$ if the following conditions are satisfied*:

$$\mu(P) = \nu(B), \tag{2.10}$$

$$L\left(\frac{1+z}{1-z}\right) \text{ is bounded away from zero on } \Lambda, \tag{2.11}$$

where Λ denotes the closed support of the singular measure of B_s.

Before proving the theorem we introduce some notation and prove a lemma. Let $\Omega \subset \mathbf{C}$ be symmetric with respect to the real axis and define

$$\nu_\Omega(P) := \inf_{Q \in \tilde{H}^\infty(\Omega)} \|W(1 - PQ)\|_{\infty,\Omega}$$

where $\|f(s)\|_{\infty,\Omega} := \sup\{|f(s)|: s \in \Omega\}$.

LEMMA 3. *If $P \in \tilde{H}^\infty(\Omega)$ is bounded away from zero on $\partial\Omega$ then $\nu_\Omega(\cdot)$ is continuous at P.*

Proof. Let $\|P_i - P\|_{\infty,\Omega} \to 0$ as $i \to \infty$. As in Lemma 1 we have $\limsup \nu_\Omega(P_i) \leqq \nu_\Omega(P)$. Write

$$0 < \delta := \inf_{s \in \partial\Omega} |P(s)|$$

and choose n such that $\|P_i - P\|_{\infty,\Omega} < \delta/2$ for all $i \geqq n$. Next take any $Q \in \tilde{H}^\infty(\Omega)$ such that there is an $m \geqq n$ with $\|W(1 - P_m Q)\|_{\infty,\Omega} \leqq \nu_\Omega(P) + 1$. Then $|P_m(s)| > \delta/2$ for $s \in \partial\Omega$

and we deduce that

$$\delta/2\|WQ\|_{\infty,\Omega} \leqq \|WP_mQ\|_{\infty,\Omega}$$
$$\leqq \|W\|_{\infty,\Omega} + \nu_\Omega(P) + 1 =: k$$

i.e., $\|WQ\|_{\infty,\Omega} \leqq 2k/\delta$. Now for any $i \geqq n$

$$\|W(1-P_iQ)\|_{\infty,\Omega} \geqq \|W(1-PQ)\|_{\infty,\Omega} - \|P-P_i\|_\infty \frac{2k}{\delta}$$
$$\geqq \nu_\Omega(P) - \|P-P_i\|_\infty \frac{2k}{\delta}.$$

It thus follows that $\liminf \nu_\Omega(P_i) \geqq \nu_\Omega(P)$ and so $\nu_\Omega(P_i) \to \nu_\Omega(P)$, completing the proof. □

Proof of Theorem 1. Let $P_i \in \tilde{H}^\infty$ be such that $\|P_i - P\|_\infty \to 0$ as $i \to \infty$. We wish to show that $\mu(P_i) \to \mu(P)$ as $i \to \infty$. First we transform to the unit disc using the substitution $z = (s-1)/(s+1)$. Now let $\Delta \subset \partial D$ be the set of points where L is not bounded away from zero. By assumption $\Delta \cap \Lambda = \phi$. Next we construct a sequence of regions $D_1 \subset D_2 \cdots \subset D$ such that P is bounded away from zero on ∂D_j and

$$\bigcup_{j=1}^{\infty} D_j = D - \Delta.$$

From Lemma 3 we have

$$\nu_{D_j}(P_i) \to \nu_{D_j}(P) \text{ as } i \to \infty \tag{2.12}$$

for all j. In a straightforward way we obtain

$$\mu(P_i) \geqq \inf_{Q \in H^\infty(D)} \|W(1-P_iQ)\|_\infty$$
$$\geqq \inf_{Q \in H^\infty(D)} \|W(1-P_iQ)\|_{\infty,D_j}$$
$$\geqq \inf_{Q \in H^\infty(D_j)} \|W(1-P_iQ)\|_{\infty,D_j} = \nu_{D_j}(P_i)$$

for any j. Thus, combining with (2.12) we get

$$\liminf \mu(P_i) \geqq \nu_{D_j}(P) \tag{2.13}$$

for any j. Now let B_k be the inner function obtained from B by dividing out all zeros of B outside D_k. From (2.13) we obviously have

$$\liminf \mu(P_i) \geqq \nu_{D_j}(B_k) \tag{2.14}$$

for any j and k. Our next step is to establish that

$$\sup_j \mu_{D_j}(B_k) = \nu(B_k). \tag{2.15}$$

Suppose to the contrary that we can find a sequence $S_j = W - B_kQ_j$ where $Q_j \in \tilde{H}^\infty(D_j)$ and $\delta > 0$ such that $\|S_j\|_{\infty,D_j} \leqq \nu(B_k) - \delta$ for all j. Then in D_n, the sequence $S_n, S_{n+1}, \cdots$ forms a normal family (see [10, p. 300]). Thus we can find a subsequence converging uniformly on compact subsets of D_n to some $S \in \tilde{H}^\infty(D_n)$. But S can be continued analytically to D_{n+1} since the subsequence is a normal family in D_{n+1}. In fact S can be continued analytically to D and moreover $\|S\|_\infty \leqq \nu(B_k) - \delta$. But we also have that

a subsequence of $Q_j = (W - S_j)B_k^{-1}$ converges uniformly to $Q := (W - S)B_k^{-1}$ on any compact subset of D. Thus $Q \in H(D)$. Moreover

$$\|Q_j\|_{\infty, D_j} \leqq \frac{(\|W\|_\infty + \nu(B_k) - \delta)}{\inf_{z \in \partial D_j} |B_k|}.$$

Since B_k is analytic on Δ we have $\inf_{z \in \partial D_j} |B_k| \to 1$ as $j \to \infty$. Thus $|Q(z) \leqq \|W\|_\infty + \nu(B_k) - \delta$ for all $z \in D$ from which we conclude that $Q \in \tilde{H}^\infty$. This means that $\nu(B_k) \leqq \|S\|_\infty$, which is a contradiction. Hence (2.15) holds and together with (2.14) we get

$$\liminf \mu(P_i) \geqq \nu(B_k) \tag{2.16}$$

for all k. The final step of the proof is to show

$$\sup_k \mu(B_k) = \nu(B). \tag{2.17}$$

We proceed in a similar way to the above. Fist we can find a sequence $S_k = W - B_k Q_k$ where $\|S_k\|_\infty = \nu(B_k) \leqq \nu(B)$, and $Q_k \in \tilde{H}^\infty$. Again the S_k form a normal family in D and there exists an $S \in \tilde{H}^\infty$ with $\|S\|_\infty \leqq \sup_k \nu(B_k) \leqq \nu(B)$ such that (without loss of generality) the sequence S_k converges to S uniformly on compact subsets of D. But $(W - S_k)B_l^{-1}$ converges in the same way to $(W - S)B_l^{-1}$ for any fixed l. Since $(W - S_k)B_l^{-1} \in \tilde{H}^\infty$ with norm less than or equal to $\|W\|_\infty + \nu(B)$ then $(W - S)B_l^{-1} \in \tilde{H}^\infty$. But from (2.11) the singular inner part of B divides B_l for all sufficiently large l. This means that $(W - S)B^{-1} \in \tilde{H}^\infty(D)$. Hence we have $\nu(B) \leqq \|S\|_\infty$ from which (2.17) follows.

Combining (2.16), (2.17), and (2.10) gives

$$\liminf \mu(P_i) \geqq \mu(P)$$

which together with Lemma 1 shows that $\mu(P_i) \to \mu(P)$ as $i \to \infty$. □

To apply Theorem 1 we need to be able to check conditions (2.10) and (2.11) for a given P and W. Condition (2.11) is straightforward. Condition (2.10) can be checked using the following theorem (see Corollary 1) which is a generalization of results of Zames and Francis [17] and Flamm [3].

THEOREM 2. *Let $W(s) \in \tilde{H}^\infty$ be a rational outer function and write $P(s) = B(s)L(s)$ where $B(s)$ is inner and $L(s)$ is outer. Assume that $B(s)$ has no essential singularities on $j\mathbb{R}$ and that $L(s)$ is bounded away from zero on $j\mathbb{R} \cup \{\infty\}$ except at finitely many zeros z_i. Then*

$$\mu(P) = \max\{\nu(B), |W(z_i)|\}. \tag{2.18}$$

Proof. If B is rational the result is straightforward and the methods of [17] apply immediately. So assume B has an essential singularity at ∞. Then $|W(\infty)| \leqq \mu(B)$ since $|W(\infty)|$ is the essential spectral radius of the Hankel H_{WB^*} which is bounded above by $\|H_{WB^*}\| = \mu(B)$. It is also clear that $|W(z_i)| \leqq \mu(P)$ for each i. Thus $\mu(P) \geqq \max\{\nu(B), |W(z_i)|\} =: \lambda$. We first show that

$$\lambda \geqq \inf_{\substack{Q \in R\tilde{H}^\infty \\ Q(\infty) = 0}} \|W - BQ\|_\infty. \tag{2.19}$$

Consider a $Q \in \tilde{H}^\infty$ such that $\|W - BQ\|_\infty = \mu(B) \leqq \lambda$. Following Flamm [3] we let $J_n = [1/(s+1)]^{1/n} \in \tilde{H}^\infty$ (with $J_n(0) = 1$). Then

$$W - BQJ_n = W(1 - J_n) + J_n(W - BQ). \tag{2.20}$$

Since $|W(\infty)| \leqq \lambda$, for any $\varepsilon > 0$ we can find an ω_0 such that $|W(j\omega)| < \lambda + \varepsilon$ for $\omega \geqq \omega_0$. Now for $\omega \in [0, \omega_0]$ it follows from (2.20) that

$$|(W - BQJ_n)(j\omega)| \leqq \|W\|_\infty |1 - J_n(j\omega_0)| + \lambda. \tag{2.21}$$

But $|1 - J_n|$ tends to zero uniformly on $[0, \omega_0]$ so the upper bound in (2.21) tends to λ as $n \to \infty$. Further, for $\omega \in [\omega_0, \infty)$ (2.20) gives

$$\begin{aligned} |(W - BQJ_n)(j\omega)| &\leqq (\lambda + \varepsilon)|1 - J_n(j\omega)| + \lambda |J_n(j\omega)| \\ &\leqq (\lambda + \varepsilon)(1 - |J_n(j\omega)|) + \lambda |J_n(j\omega)| \\ &\quad + (\lambda + \varepsilon)|J_n(j\omega)| - |J_n(j\omega)| \\ &\leqq (\lambda + \varepsilon)\left(1 + \frac{\pi}{2n}|J_n(j\omega)|\right). \end{aligned}$$

Thus,

$$\limsup_{n \to \infty} \|W - BQJ_n\|_\infty \leqq \lambda. \tag{2.22}$$

Now for any n we can find an N such that $|Q(s)J_n(s)| < \delta/2$ for $|s| > N$, $s \in \mathbf{C}_+$ and so

$$|(QJ_n)(j\omega + \varepsilon) - (WB^*)(j\omega)| \leqq |(QJ_n)(j\omega) - (WB^*)(j\omega)| + \delta \tag{2.23}$$

for $|\omega| > N$ and any ε. Also it is clear that

$$|(QJ_n)(j\omega + \varepsilon) - (WB^*)(j\omega + \varepsilon)| \leqq \|W - BQJ_n\|_\infty$$

for $\omega \in [-N, N]$ and any ε. Since WB^* is analytic on $j\mathbb{R}$ then $(WB^*)(j\omega + \varepsilon)$ tends uniformly to $(WB^*)(j\omega)$ on $[-N, N]$, and we deduce

$$\limsup_{\varepsilon \to 0} \sup_{\omega \in [-N, N]} |(QJ_n)(j\omega + \varepsilon) - (WB^*)(j\omega)| \leqq \|W - BQJ_n\|_\infty. \tag{2.24}$$

(2.23) and (2.24) together give

$$\limsup_{\varepsilon \to 0} \|W(s) - B(s)(QJ_n)(s + \varepsilon)\|_\infty \leqq \|W - BQJ_n\|. \tag{2.25}$$

But $(QJ_n)(s + \varepsilon)$ is continuous on $\tilde{\mathbf{C}}_+$ for any ε and can thus be approximated uniformly by rational functions (Mergelyan's Theorem; see [10]). Further we can ensure that the approximants vanish at ∞. This together with (2.25) and (2.22) establishes (2.19).

We next show that (2.19) holds with the additional constraint that $Q(z_i) = 0$ for all i. There are several approaches here (e.g., [17] where $X = W - BQ$ is modified multiplicatively). The above approach for ∞ can also be used. For simplicity assume $z_1 = 0$ and let $J_n = [s/(s+1)]^{1/n}$. Then since $|J_n(j\omega)| - |J_n(j\omega)|$ tends to zero uniformly we again have

$$\limsup_{n \to \infty} \|W - BQJ_n\|_\infty \leqq \|W - BQ\|_\infty$$

for any Q. We can then approximate QJ_n uniformly with rational functions which vanish at both 0 and infinity. Each of the z_i can be dealt with similarly.

Next we observe that (2.19) holds with Q constrained to have zeros at the z_i and ∞ of arbitrary multiplicity. This follows since, for example, if $Q(0) = 0$, then $[s/(s+\varepsilon)]^l Q$ tends to Q uniformly as $\varepsilon \to 0$. We have thus shown that

$$\begin{aligned} \lambda &\geqq \inf_{\substack{Q \in R\tilde{H}^\infty \\ (s+1)Q/WL \in \tilde{H}^\infty}} \|W - BQ\|_\infty \\ &= \inf_{\substack{Q' \in \tilde{H}^\infty \\ Q'L \in \tilde{H}^\infty \cap \mathbb{R}_{sp}(s)}} \|W(1 - PQ')\|_\infty \geqq \mu(P). \end{aligned}$$

The last inequality follows since $Q'L \in \tilde{H}^\infty \cap \mathbb{R}_{sp}(s)$ implies $(1-BLQ')^{-1} \in \tilde{H}^\infty + \mathbb{R}_{pr}(s)$. Together with the obvious inequality $\mu(P) \geqq \lambda$ this completes the proof. □

COROLLARY 1. *For $W(s)$ and $P(s)$ satisfying the conditions of Theorem 2 we have $\mu(P) = \nu(P)$ if and only if $|W(z_i)| \leqq \nu(B)$ at all zeros z_i of $L(s)$ on $j\mathbb{R} \cup \{\infty\}$.*

If we write $\mu(P, W) := \mu(P) = \inf \| W(1+PF)^{-1} \|_\infty$ we can deduce the following.

COROLLARY 2. *If $P(s)$ and $W(s)$ satisfy the conditions of Theorem 2 and W_k is a sequence of rational outer functions such that $\| W_k - W \|_\infty \to 0$ as $k \to \infty$, then $\mu(P, W_k) \to \mu(P, W)$.*

Proof. Certainly $|W_k(z_i)| \to |W(z_i)|$ as $k \to \infty$. Also $\| W_k B^* - WB^* \|_\infty \to 0$ as $k \to \infty$ and so $d(W_k B^*, H^\infty) \to d(WB^*, H^\infty)$. The conclusion now follows from (2.18). □

3. H^p weighted sensitivity minimization. In this section we consider the problem:

$$\hat{\mu}(P) := \inf_{F \text{ stablz.}} \| W(1+PF)^{-1} \|_p \tag{3.1}$$

where $1 \leqq p < \infty$, $P(s) \in \tilde{H}^\infty$ and $F(s) \in \tilde{H}^\infty + \mathbb{R}_{pr}(s)$. We take $W(s)$ to be a rational outer function such that $W \in (s+1)^{-l} \tilde{H}^\infty$ where l is the least integer with $l > 1/p$ (so that $W \in \tilde{H}^p$). We will again consider the question: when does $\| P_i - P \|_\infty \to 0$ as $i \to \infty$ imply that $\hat{\mu}(P_i) \to \hat{\mu}(P)$? As for the case $p = \infty$ it is easy to show $\hat{\mu}(\cdot)$ is upper semicontinuous. Also we have in an obvious way that

$$\begin{aligned} \hat{\mu}(P) &= \inf_{\substack{Q \in \tilde{H}^\infty \\ Q/(1-PQ) \in \tilde{H}^\infty + \mathbb{R}_{pr}(s)}} \| W(1-PQ) \|_p \\ &\geqq \inf_{Q \in \tilde{H}^\infty} \| W(1-PQ) \|_p \\ &\geqq \inf_{Q \in \tilde{H}^\infty} \| W(1-BQ) \|_p =: \hat{\nu}(B). \end{aligned} \tag{3.2}$$

In fact we can prove the following result.

THEOREM 3. *For any $P(s)$ satisfying the conditions of Theorem 2*

$$\hat{\mu}(P) = \hat{\nu}(B) = \inf_{Q \in \tilde{H}^p} \| WB^* - Q \|_p = \inf_{Q \in (s+1)^{-N} \tilde{H}^\infty_-} \| WB^* - Q \|_p \tag{3.3}$$

(where $N > 0$ can be chosen arbitrarily).

Proof. Since $WQ \in \tilde{H}^p$ for any $Q \in \tilde{H}^\infty$ it immediately follows that

$$\hat{\nu}(B) \geqq \inf_{Q \in \tilde{H}^p} \| WB^* - Q) \|_p. \tag{3.4}$$

But $(s+1)^{-N} \tilde{H}^\infty_-$ is dense in $\tilde{H}^p$ for any N (see [6, pp. 59-60]), which shows the third equality in (3.3). Now given any $Q \in (s+1)^{-N} \tilde{H}^\infty_-$ we have $\| [s/(s+\varepsilon)]^r Q - Q \|_p \to 0$ as $\varepsilon \to 0$ by Lebesgue's dominated convergence theorem [10]. Thus

$$\begin{aligned} \inf_{Q \in (s+1)^{-N} \tilde{H}^\infty_-} \| WB^* - Q \|_p &= \inf_{\substack{Q \in (s+1)^{-N} \tilde{H}^\infty_- \\ Q/WL \in \tilde{H}^\infty}} \| WB^* - Q \|_p \\ &= \inf_{\substack{Q' \in \tilde{H}^\infty \\ Q'L \in (s+1)^{-N} \tilde{H}^\infty_-}} \| W(1-BLQ') \|_p \geqq \hat{\mu}(P), \end{aligned}$$

which together with (3.2) and (3.4) completes the proof. □

Theorem 3 shows that, for $1 \leqq p < \infty$, Example 1 is no longer ill-posed. More generally we have the following corollary.

COROLLARY. *If $P(s) \in \mathbb{R}_{pr}(s)$ then $\| P_i - P \|_\infty \to 0$ implies $\hat{\mu}(P_i) \to \hat{\mu}(P)$.*

Proof. Write $P = BL$ where B is inner and L is outer. Now Rouché's Theorem shows that the $\mathbf{C}_+$ zeros of P_i tend to those of P as $i \to \infty$. Thus we can write $P_i = B_i L_i$

where B_i is a finite Blaschke product (L_i not necessarily outer) and $\|B_i - B\|_\infty \to 0$ as $i \to \infty$. Now $\hat{\mu}(P_i) \geqq \hat{\mu}(B_i)$ and, since $\|WB - WB_i\|_p \to 0$ as $i \to \infty$, $\hat{\mu}(B_i) \to \hat{\mu}(B) = \hat{\mu}(P)$. Thus $\liminf \hat{\mu}(P_i) \leqq \hat{\mu}(P)$ from which the result follows by the upper semicontinuity of $\hat{\mu}(\cdot)$. □

However, the ill-posed in Example 2 remains, as we now show.

Example 2 (continued). Assume $1 < p < \infty$, $P = e^{-s}/(s+1)$ and $W = 1/(s+1)$. Then

$$\hat{\mu}(P) = \inf_{Q \in \tilde{H}^p} \left\| \frac{e^s}{s+1} - Q \right\|_p.$$

Since $e^s/(s+1) \notin \tilde{H}^p$, $\hat{\mu}(P) > 0$. In fact, for $p = 2$ we can explicitly compute $\hat{\mu}(P)$ as follows. For any $Q \in \tilde{H}^2$

$$\left\| \frac{e^s}{s+1} - Q \right\|_2^2 = \left\| \frac{e^s - e^{-1}}{s+1} \right\|_2^2 + \left\| \frac{e^{-1}}{s+1} - Q \right\|_2^2$$
$$\geqq (1 + e^{-2})\pi = \hat{\mu}(P)^2.$$

But $\|P_n - P\|_\infty \to 0$ as $n \to \infty$ where $P_n(s)$ is given by (2.5), and moreover, from Theorem 3, $\hat{\mu}(P_n) = 0$ for all n. Thus $\hat{\mu}(\cdot)$ is not continuous at P. □

Once again we can establish a general condition for continuity which parallels Theorem 1.

THEOREM 4. *For any $P \in \tilde{H}^\infty$, $\hat{\mu}(\cdot)$ is continuous at P if* (2.11) *holds.*

Proof (sketch). It is convenient to transform to the disc. This gives

$$\hat{\mu}(P) = \inf_{Q \in \tilde{H}^\infty(D)} \pi^{-1/p} \|(z-1)^{-2/p} \tilde{W}(1 - \tilde{P}\tilde{Q})\|_p$$
$$= \inf_{Q \in \tilde{H}^p(D)} \pi^{-1/p} \|\tilde{W}'(1 - \tilde{P}\tilde{Q})\|_p$$

where $\tilde{W}' = (z-1)^{-2/p}\tilde{W}$. As before, Lemma 3 holds (this time we have to bound $\|\tilde{W}'\tilde{Q}\|_p$). The proof now follows that of Theorem 1 with some modifications. We note that $\|f\|_p \geqq \|f\|_{p,D_j}$ follows from the subharmonicity of $|f|^p$ (see [10]). The two normal family arguments also go through if we work with compact subsets of the disc. □

4. Achieving continuity of performance measure. So far we have considered the problem of sensitivity minimization and have shown that optimal sensitivity is not a continuous function on H^∞. One might expect that this problem would not arise if other terms were included in the objective function, e.g., if a trade-off were sought between sensitivity and complementary sensitivity. However the problem can still occur in the latter case. Consider the following example.

Example 3. Let $P(s) = s/(s+1)$, set

$$\mu(P) = \inf_{F \text{ stablz.}} (\|W_1 PF(1+PF)^{-1}\|_\infty + \|W_2(1+PF)^{-1}\|_\infty)$$
$$= \inf_{Q \in H^\infty} (\|W_1 PQ\|_\infty + \|W_2(1-PQ)\|_\infty)$$

and suppose $\|W_1\|_\infty < |W_2(0)|$. Certainly $\mu(P) \geqq |W_2(0)|$ by considering the second term only. But $\|P_\varepsilon - P\|_\infty \to 0$ as $\varepsilon \to 0$ where $P_\varepsilon = (s+\varepsilon)/(s+1)$, and furthermore $\mu(P_\varepsilon) \leqq \|W_1\|_\infty$ (just by taking $PQ = 1$). Thus we have discontinuity at $P(s)$.

On the other hand, certain optimization problems do behave continuously. By considering the proof of Lemma 3 we can discover a way to force continuity of the performance measure. Consider a problem of the following type:

$$\mu(P) = \inf_{Q \in \mathscr{S}} m(P, Q) \tag{4.1}$$

where $\mathcal{S} \subset \tilde{H}^\infty$ and $m(\cdot,\cdot)$ is a nonnegative function on $\tilde{A}^\infty \times \mathcal{S}$. We remark that many frequency domain optimization problems take the form (4.1). In the theorem below we will show that $\mu(\cdot)$ *is continuous on* $\tilde{A}^\infty$ *at* P *if in some neighbourhood of* P *there are minimizing sequences* Q_i *which are uniformly bounded.*

THEOREM 5. *Let* $P_0 \in \tilde{A}^\infty$ *and suppose for any* $P \in \mathcal{B} := \{P \in \tilde{A}^\infty : \|P - P_0\|_\infty < \varepsilon\}$ *(some* $\varepsilon > 0$*) and* $Q \in \mathcal{S}$ *we can find* $c > 0$ *and* $W \in \tilde{H}^\infty$ *such that*

$$|m(P, Q) - m(P_0, Q)| \leqq c\|P - P_0\|_\infty \|WQ\|_p \tag{4.2}$$

whenever $WQ \in \tilde{H}^p$. *Then if there exists* $M > 0$ *such that for any* $P \in \mathcal{B}$ *we can find a sequence* $Q_i \in \mathcal{S}$ *with* $\|WQ_i\|_p < M$ *and* $m(P, Q_i) \to \mu(P)$, *then* $\mu(\cdot)$ *is continuous at* P_0.

Proof. Let $\|P_n - P_0\|_\infty \to 0$ as $n \to \infty$. Then $m(P_n, Q) \to m(P_0, Q)$ shows that $\limsup \mu(P_n) \leqq \mu(P_0)$. But for any n, we deduce from (4.2) that $\mu(P_n) \geqq \mu(P_0) - cM\|P_0 - P_n\|_\infty$. Hence $\mu(P_n) \to \mu(P)$. □

It is now easy to write down problems which are well-posed in the sense that $\mu(\cdot)$ is continuous on $\tilde{A}^\infty$.

COROLLARY. *Define* $\mu(P)$ *as in* (4.1). *Then* $\mu(\cdot)$ *is continuous on* $\tilde{A}^\infty$ *for the following choices of* $\mathcal{S}$ *and* $m(\cdot,\cdot)$

$$\mathcal{S} = \tilde{H}^\infty,\ m(P, Q) = \left\| \begin{matrix} W_1 Q \\ W_2(1 - PQ) \end{matrix} \right\|_\infty \tag{4.3}$$

$$\mathcal{S} = \{Q \in \tilde{H}^\infty : \|Q\|_\infty < M\},\ m(P, Q) = \|W(1 - PQ)\|_p \tag{4.4}$$

for any $1 \leqq p \leqq \infty$, W_1, $W_1^{-1} \in R\tilde{H}^\infty$ *and* $W \in \tilde{H}^p \cap \mathbb{R}_{sp}(s)$.

For $p = \infty$, problem (4.1+3) represents a desirable choice of design problem since there are procedures for solving this exactly in the rational case (e.g., [5], [8], and [13]). Since we are considering $P \in \tilde{A}^\infty$, it follows as in Theorem 2 that problem (4.1+3) is precisely:

$$\mu(P) = \inf_{F \text{ stablz.}} \left\| \begin{matrix} W_1 F(1 + PF)^{-1} \\ W_2(1 + PF)^{-1} \end{matrix} \right\|_\infty \tag{4.5}$$

(and similarly for (4.1+4)).

Problem (4.1+4) can be viewed as H^p weighted sensitivity minimization with a constraint on $\|Q\|_\infty$. Such a bound does make good physical sense. In the first place, Q as defined in (2.2) is the transfer function from $x_2 \to e_1$. So (4.1+4) represents a disturbance attenuation problem *subject to a disturbance to plant input power limitation.* Second, $\|Q\|_\infty < M$ guarantees a measure of robustness. If $F = Q/(1 - P_0 Q)$ stabilizes P_0 then F stabilizes any P with $\|P - P_0\|_\infty < 1/M$ (since $F(1 + PF)^{-1} = Q(1 - (P - P_0)Q)^{-1}$).

5. The robust sensitivity minimization problem (RSMP). In light of the observation in § 4, that weighted sensitivity minimization subject to a constraint on $\|Q\|_\infty$ is well-posed, it is natural to turn our attention to the problem of robust weighted sensitivity minimization. For stable plants this can be defined as follows. Let $P_0(s) \in \tilde{H}^\infty$ and define

$$\mathcal{B}(\alpha) = \{P(s) \in \tilde{H}^\infty : |(P - P_0)(j\omega)| \leqq \alpha r(j\omega)\} \tag{5.1}$$

where $r(j\omega)$ is a nonnegative L^∞-function. Then we are required to find:

$$\mu(\alpha) = \inf_{\substack{F \text{ which stabilize} \\ \text{all } P \in \mathcal{B}(\alpha)}} \sup_{P \in \mathcal{B}(\alpha)} \|W(1 + PF)^{-1}\|_\infty \tag{5.2}$$

for a given rational outer function W. The question arises whether $\mu(\alpha)$ and the associated optimal control are continuous functions of α. The answer is no, as we now show.

Example 4. Let $P_0(s) = 1/(s+2)$, $r(j\omega) = 1/|1+j\omega|$ and $W(s) = 1/(s+1)$. We claim that $\mu(\alpha)$ is discontinuous and in fact

$$\mu(\alpha) = \begin{cases} 0 & \text{for } \alpha < \frac{1}{2} \\ \geqq 1 & \text{for } \alpha \geqq \frac{1}{2}. \end{cases} \tag{5.3}$$

Clearly if $\alpha \geqq \frac{1}{2}$ then

$$P_1 = \frac{1}{s+2} - \frac{1}{2}\frac{1}{s+1} = \frac{s}{2(s+1)(s+2)} \in \mathcal{B}(\alpha).$$

Therefore for any stabilizing F, $\|W(1+P_1F)^{-1}\|_\infty \geqq |W(0)| = 1$. Thus $\mu(\alpha) \geqq 1$. Now suppose $\alpha < \frac{1}{2}$. We first claim that $F = k$ (for any $k > 0$) is a stabilizing feedback for all $P \in \mathcal{B}(\alpha)$. Write

$$(1+Pk)^{-1} = \left(1+\frac{k}{s+2}\right)^{-1}\left(1+\alpha k\Delta\left(1+\frac{k}{s+2}\right)^{-1}\right)^{-1} \tag{5.4}$$

where $\Delta(s) \in \tilde{H}^\infty$ and $|\Delta(j\omega)| \leqq 1/|1+j\omega|$. Then

$$\left\| \alpha k\Delta\left(1+\frac{k}{s+2}\right)^{-1} \right\|_\infty \leqq \left\| \frac{\alpha k(s+2)}{(s+1)(s+2+k)} \right\|_\infty \leqq 2\alpha \left\| \frac{k}{s+2+k} \right\|_\infty \leqq 2\alpha < 1.$$

Thus from (5.4) we see that $(1+Pk)^{-1} \in \tilde{H}^\infty$. Furthermore

$$\sup_{P\in\mathcal{B}(\alpha)} \|W(1+Pk)^{-1}\|_\infty \leqq \left\| \frac{1}{s+1}\left(1+\frac{k}{s+2}\right)^{-1} \right\|_\infty \cdot \frac{1}{1-2\alpha}. \tag{5.5}$$

But the right-hand side of (5.5) tends to zero as $k \to \infty$. Thus $\mu(\alpha) = 0$.

Owing to the difficulty of solving the RSMP in general a complete analysis of the well-posedness issue here seems rather hard. In the above example

$$\lim_{k\to\infty} \sup_{P\in\mathcal{B}(\alpha)} \|k(1+Pk)^{-1}\|_\infty$$

tends to infinity as $\alpha \uparrow \frac{1}{2}$. This forces a new control strategy to be adopted for $\alpha \geqq \frac{1}{2}$. Once again it turns out that if we keep

$$\sup_{P\in\mathcal{B}(\alpha)} \|F(1+PF)^{-1}\|_\infty$$

bounded then we can ensure continuity. This is the essence of the following result.

THEOREM 6. *Let $\mu(\alpha)$ be defined by*

$$\mu(\alpha) = \inf_{\substack{F \text{ which stabilize} \\ \text{all } P\in\mathcal{B}(\alpha)}} \sup_{P\in\mathcal{B}(\alpha)} m(P, F) \tag{5.6}$$

where

$$m(P, F) = \|W_1(1+PF)^{-1}\|_\infty + \|W_2F(1+PF)^{-1}\|_\infty \tag{5.7}$$

and $W_2W_2^{-1} \in \tilde{H}^\infty$. Then $\mu(\alpha)$ is continuous on $[0, \infty)$.

Proof. It is immediately clear that $\mu(\alpha_1) \leqq \mu(\alpha_2)$ for $\alpha_1 \leqq \alpha_2$. We now show that $\mu(\alpha)$ is upper semicontinuous. Let $\alpha_i \downarrow \alpha_0$. Then $\lim \mu(\alpha_i) \geqq \mu(\alpha_0)$. Now take any F which stabilizes all $P \in \mathcal{B}(\alpha_0)$. Then F stabilizes all $P \in \mathcal{B}(\alpha_i)$ for i sufficiently large since $\mathcal{B}(\alpha)$ is closed. Furthermore, defining

$$\eta(F, \alpha) := \sup_{P\in\mathcal{B}(\alpha)} m(P, F)$$

we see that $\eta(F, \alpha_i) \to \eta(F, \alpha_0)$ as $i \to \infty$, because $m(P, F)$ depends continuously on P. Thus $\lim \mu(\alpha_i) = \mu(\alpha_0)$.

Now the form of (5.7) allows us to show that $\mu(\alpha)$ is lower semicontinuous. First note that if $\hat{\alpha} \leqq \alpha_0$ and F stabilizes all $P \in \mathcal{B}(\hat{\alpha})$ with $\eta(F, \hat{\alpha}) \leqq \mu(\alpha_0)+1$ then

$$\sup_{P \in \mathcal{B}(\hat{\alpha})} \|F(1+PF)^{-1}\|_\infty \leqq \frac{\mu(\alpha_0)+1}{\|W_2^{-1}\|_\infty} =: c. \tag{5.8}$$

We now claim that, for the same F,

$$\sup_{P \in \mathcal{B}(\hat{\alpha}+d)} \|F(1+PF)^{-1}\|_\infty \leqq 2c \tag{5.9}$$

where $d := (2c \sup r(j\omega))^{-1}$. To see this take any $P \in \mathcal{B}(\hat{\alpha}+d)$ and write $P = P_1 + \Delta$ where $P_1 = P_0 + (P-P_0)\hat{\alpha}/(\hat{\alpha}+d) \in \mathcal{B}(\hat{\alpha})$ and $\Delta = (P-P_0)d/(\hat{\alpha}+d) \in \tilde{H}^\infty$, $\|\Delta\|_\infty \leqq (2c)^{-1}$. Then

$$F(1+(P_1+\Delta)F)^{-1} = F(1+P_1F)^{-1}(1+\Delta F(1+P_1F)^{-1})^{-1}$$

and it follows that $\|F(1+PF)^{-1}\|_\infty \leqq c(1-1/2)^{-1}$, which establishes (5.9). Now consider a sequence $\alpha_i \uparrow \alpha_0$ such that $\alpha_i > \alpha_0 - d$ and take a sequence of F_i (stabilizing all $P \in \mathcal{B}(\alpha_i)$) such that $\eta(F_i, \alpha_i) \leqq \mu(\alpha_i) + 1/(i+1)$. Then each F_i stabilizes all $P \in \mathcal{B}(\alpha_0)$ from (5.9). But for any $P_1 \in \mathcal{B}(\alpha_i)$, $P_2 \in \mathcal{B}(\alpha_0)$ and $F = F_i$,

$$\begin{aligned} |m(P_1, F) - m(P_2, F)| &\leqq \|W_1(1+P_1F)^{-1}(P_2-P_1)F(1+P_2F)^{-1}\|_\infty \\ &\quad + \|W_2F(1+P_1F)^{-1}(P_2-P_1)F(1+P_2F)^{-1}\|_\infty \\ &\leqq A\|P_2-P_1\|_\infty \end{aligned}$$

for some positive constant A. This means that

$$\mu(\alpha_0) \leqq \mu(\alpha_i) + 1/(i+1) + B(\alpha_0 - \alpha_i)$$

for some positive constant B, and we conclude, since $\mu(\alpha_i) \leqq \mu(\alpha_0)$ for all i, that $\mu(\alpha_i) \to \mu(\alpha_0)$ as $i \to \infty$. We have thus shown lower semicontinuity, which completes the proof. □

Remark. The theorem holds for any $m(P, F)$ which guarantees a bound of the form (5.8).

6. Unstable plants—continuity of performance measure. We now consider the class $\mathcal{P} = \tilde{A}^\infty + \mathbb{R}_{pr}(s)$. If $P \in \mathcal{P}$ then there exists $A \in \tilde{H}^\infty$, $B \in \tilde{H}^\infty$, and X, $Y \in \tilde{H}^\infty$ with

$$P = \frac{A}{B}, \qquad AX + BY = 1. \tag{6.1}$$

We define a topology on plants $P \in \mathcal{P}$ as follows ([2], [14]). We say $P_n \to P$ if we can find sequences A_n, $B_n \in \tilde{H}^\infty$ such that $P_n = A_n/B_n$ and $\|A_n - A\|_\infty + \|B_n - B\|_\infty \to 0$. It follows that for sufficiently large n there exist X_n, $Y_n \in \tilde{H}^\infty$ such that $A_nX_n + B_nY_n = 1$ and $\|X_n - X\|_\infty + \|Y_n - Y\|_\infty \to 0$ as $n \to \infty$.

It is easy to see that keeping $\|F(1+PF)^{-1}\|_\infty$ bounded is no longer sufficient to ensure continuity of a general optimization problem.

Example 5. Let $P(s) = (s+1)/s$ and consider

$$\mu(P) = \inf_{F \text{ stablz.}} (\|F(1+PF)^{-1}\|_\infty + \|WPF(1+PF)^{-1}\|_\infty)$$

where $W(0) \neq 0$. Considering the second term at $s=0$ shows $\mu(P) \geqq |W(0)|$. But $P_\varepsilon = (s+1)/(s+\varepsilon) \to P$ as $\varepsilon \to 0$ and $\mu(P_\varepsilon) = 0$ since $F = 0$ is a stabilizing feedback. Thus $\mu(\cdot)$ is not continuous at P.

To motivate a sufficient condition for continuity of the performance measure we recall that the feedback system of Fig. 1 is stable if and only if

$$F(1+PF)^{-1} \in \tilde{H}^{\infty} \quad \text{and} \quad P(1+PF)^{-1} \in \tilde{H}^{\infty}. \tag{6.2}$$

We would therefore expect that keeping the norms of both transfer functions in (6.2) bounded would force continuity. This is indeed the case for a very general class of performance criteria. To see this we observe that any stabilizing feedback for P can be written in the form $F = (X + BQ)/(Y - AQ)$ for some $Q \in \tilde{H}^{\infty}$. If we substitute for F we find that any closed loop transfer function is an affine function of Q:

$$T_1 := (1+PF)^{-1} = BY - ABQ, \tag{6.3a}$$

$$T_2 := F(1+PF)^{-1} = BX + B^2Q, \tag{6.3b}$$

$$T_3 := P(1+PF)^{-1} = AY - A^2Q, \tag{6.3c}$$

$$T_4 := PF(1+PF)^{-1} = AX + ABQ. \tag{6.3d}$$

Now define

$$\mu(P) = \inf_{F \text{ stablz.}} m(P, F)$$

where

$$m(P, F) = \sum_{i=1}^{4} \| W_i T_i \|_{\infty} = r(Q)$$

and $W_i \in R\tilde{H}^{\infty}$ are outer functions (possibly zero). Once again, we can show as in Theorem 2 that taking the infimum over *realizable* F's is equivalent to taking the infimum of $r(Q)$ over all $Q \in \tilde{H}^{\infty}$, i.e.,

$$\mu(P) = \inf_{Q \in \tilde{H}^{\infty}} r(Q). \tag{6.4}$$

Now, using the same reasoning as in the proof of Theorem 5 we can show the following result.

THEOREM 7. *Consider any sequence $P_n \to P \in \mathcal{P}$. Then $\mu(P_n) \to \mu(P)$ if for some $k > 0$ and any n sufficiently large there exists a minimizing sequence $Q_{n,i} \in \tilde{H}^{\infty}$ (i.e., $m(P_n, F_{n,i}) \to \mu(P_n)$ as $i \to \infty$ where $F_{n,i} = (X_n + B_n Q_{n,i})/(Y_n - A_n Q_{n,i})$) with $\| Q_{n,i} \|_{\infty} \leqq k$.*

COROLLARY. *If W_2^{-1}, $W_3^{-1} \in \tilde{H}^{\infty}$ then $\mu(\cdot)$ is continuous at all $P \in \mathcal{P}$.*

Proof. Consider any $P \in \mathcal{P}$. Then we can find an ε such that for any $P_1 = A_1/B_1 \in \mathcal{P}$ with $A_1X_1 + B_1Y_1 = 1$ and

$$\| A - A_1 \|_{\infty}, \quad \| B - B_1 \|_{\infty}, \quad \| X - X_1 \|_{\infty}, \quad \| Y - Y_1 \|_{\infty} < \varepsilon,$$

there holds $\mu(P_1) \leqq \mu(P) + 1$. Now if Q_i is any minimizing sequence for P_1 such that $r(Q_i) \leqq \mu(P_1) + 1$, then we certainly have

$$\| A_1 Y_1 - A_1^2 Q_i \|_{\infty} \leqq (\mu(P) + 2) \| W_2^{-1} \|_{\infty}.$$

$$\| B_1 X_1 + B_1^2 Q_i \|_{\infty} \leqq (\mu(P) + 2) / \| W_3^{-1} \|_{\infty}.$$

We therefore deduce that

$$\sup_{s \in \mathbf{C}_+} |Q_i(s)| (|A_1(s)|^2 + |B_1(s)|^2) \leqq M$$

for some constant M. But from (6.1) we have

$$\inf_{s \in \mathbf{C}_+} (|A(s)|^2 + |B(s)|^2) \geqq \delta > 0$$

for some δ. Thus by taking ε sufficiently small we obtain $\|Q_i\|_\infty \leqq 2M/\delta$ for all i. □

7. Concluding remarks.

7.1. Choice of topology. In this paper we have defined two stable plants to be close if the *absolute* error in their frequency responses is uniformly small. Alternatively we could define a stronger topology on plants by requiring that the *relative* error in their frequency responses be uniformly small, i.e., $P_n \to P$ if

$$\sup_\omega \left| \left(\frac{P_n - P}{P} \right)(j\omega) \right| \to 0 \tag{7.1}$$

as $n \to \infty$. From (7.1) it is easy to show that $\mu(P_n) \to \mu(P)$ for H^∞ weighted sensitivity minimization (or any other problem we have considered). To verify this we need to show, as in Lemma 3, that $\liminf \mu(P_n) \geqq \mu(P)$. So take any sequence $Q_n \in \tilde{H}^\infty$ such that $\|W(1-P_nQ_n)\|_\infty \leqq \mu(P_n) + 1/n$. Then clearly $\|WPQ_n\|_\infty \leqq M$ for some positive constant M. But

$$\begin{aligned} |\,\|W(1-P_nQ_n)\|_\infty - \|W(1-PQ_n\|_\infty| &\leqq \|W(P-P_n)Q_n\|_\infty \\ &\leqq \sup_\omega \left| \left(\frac{P-P_n}{P} \right)(j\omega) \right| \|WPQ_n\|_\infty \end{aligned}$$

which implies that $\liminf \|W(1-PQ_n)\|_\infty \leqq \liminf \mu(P_n)$. Hence $\mu(P_n) \to \mu(P)$. However, to strengthen the topology in this way does not really resolve the ill-posedness problem satisfactorily from the engineering point of view since the topology is *too* strong to be relevant in most physical situations. To obtain a close approximation in the sense of (7.1) would require accurate magnitude and phase information at frequencies where the magnitude of the frequency response was arbitrarily small.

7.2. Approximation of delay systems. In Example 2 $\mu(P)$ was shown to be discontinuous at $P(s) = e^{-s}/(s+1)$ by approximating $P(s)$ with a sequence of outer functions. On the other hand, if we have a sequence $P_n = B_n/(s+1)$ such that $\|P_n - P\|_\infty \to 0$ as $n \to \infty$, and further, if B_n is inner and satisfies

$$\|W(B_n^* - e^s)\|_\infty \to 0 \tag{7.2}$$

as $n \to \infty$, then $\mu(P_n) \to \mu(P)$. Thus $\mu(P)$ could be computed by a sequence of (rational) approximants to P, given appropriate choices of B_n. Such a choice is given by $B_n = (1-s/2n)^n/(1+s/2n)^n$ in case $W(s)$ is strictly proper. If $W(\infty) \neq 0$ then (7.2) fails and it is no longer clear how to guarantee $\mu(P_n) \to \mu(P)$.

7.3. Continuous dependence of control on plant. This paper has been concerned with the continuous dependence of the infimum $\mu(P)$, defined by certain minimization problems, on the plant P. A more challenging problem is to arrange for the optimal control F_{opt} to depend continuously on P. Continuity of $\mu(P)$ is by no means sufficient for this. In the H^∞ optimal sensitivity problem it is certainly possible that a rational plant satisfies $\mu(P) = \nu(B)$ and that the infimum in (2.3) is not achieved (e.g., consider a plant which is strictly proper or has a zero on the imaginary axis). In this case the map $P \to F_{\text{opt}}$ is not even well-defined. If we relax the condition that the infimum be achieved exactly, the problem then becomes to select an appropriate F from among infinitely many which are within an arbitrary ε of optimality, and moreover to do this continuously as a function of P. A similar difficulty arises if the infimum to the design problem is not achieved uniquely (which typically occurs, for example, in multivariable H^∞ optimization problems).

Acknowledgment. I would like to thank an anonymous reviewer for some helpful remarks on the manuscript.

REFERENCES

[1] V. M. ADAMJAN, D. Z. AROV, AND M. G. KREIN, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem*, Math. USSR Sb., 15 (1971), pp. 31-73.

[2] A. K. EL-SAKKARY, *The gap metric: robustness of stabilization of feedback systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 240-247.

[3] D. S. FLAMM, *Control of delay systems for minimax sensitivity*, Ph.D. thesis, LIDS-TH-1560, Massachusetts Institute of Technology, Cambridge, 1986.

[4] C. FOIAS, A. TANNENBAUM, AND G. ZAMES, *Weighted sensitivity minimization for delay systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 763-766.

[5] B. A. FRANCIS, *A course in H_∞ control theory*, Lecture Notes in Control and Information Sciences 88, Springer-Verlag, Berlin, New York, 1987.

[6] J. B. GARNETT, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[7] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice Hall, Englewood Cliffs, NJ, 1962.

[8] H. KWAKERNAAK, *Minimax frequency domain performance and robustness optimization of linear feedback systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 994-1004.

[9] Z. NEHARI, *On bounded bilinear forms*, Ann. of Math., 65 (1957), pp. 153-162.

[10] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1974.

[11] D. SARASON, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179-203.

[12] M. C. SMITH, *Sensitivity of H^∞ optimal control problems*, Linear Circuits, Systems and Signal Processing, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North-Holland, Amsterdam, 1988, pp. 597-602.

[13] M. S. VERMA AND E. A. JONCKHEERE, *L_∞-compensation with mixed sensitivity as a broadband matching problem*, Sys. Control Lett., 4 (1984), pp. 125-129.

[14] M. VIDYASAGAR, *The graph metric for unstable plants and robustness estimates for feedback stability*, IEEE Trans. Automat. Control, 29 (1984), pp. 403-418.

[15] J. L. WALSH, *Interpolation and Approximation by Rational Functions in the Complex Domain*, Amer. Math. Soc. Colloq. Publ., vol. 20, 1960.

[16] G. ZAMES, *Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, 26 (1981), pp. 301-320.

[17] G. ZAMES AND B. A. FRANCIS, *Feedback, minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, 28 (1983), pp. 585-601.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 359-381, March 1990

006

QUADRATIC CONTROL FOR LINEAR TIME-VARYING SYSTEMS*

GIUSEPPE DA PRATO† AND AKIRA ICHIKAWA‡

Abstract. An infinite-dimensional linear time-varying system on the interval $(-\infty, \infty)$ is considered. We introduce three quadratic problems: the infinite horizon problem, and one-sided and two-sided average cost problems. A Riccati equation on $(-\infty, \infty)$ is considered first and sufficient conditions for the existence and uniqueness of a bounded solution are given. Then by dynamic programming the quadratic problems are solved. Similar problems in the stochastic case are considered.

Key words. linear quadratic control, time-varying systems

AMS(MOS) subject classifications. 49, 49C20

1. Introduction.

Consider the usual quadratic control problem:

$$y' = A(t)y + B(t)u, \qquad y(t_0) = y_0, \tag{1.1}$$

$$J(u) = \int_{t_0}^{T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt \tag{1.2}$$

where A, B, M, and N, are continuous matrices on $(-\infty, \infty)$ of appropriate dimensions and where $|\ |$ and $\langle\ ,\ \rangle$ denote, respectively, the norm and the inner product of vectors. The Riccati equation associated with this problem is the following [28]:

$$Q' + A^*Q + QA + M^*M - QBN^{-1}B^*Q = 0, \tag{1.3}$$

$$Q(T) = 0. \tag{1.4}$$

There exists a unique solution to (1.3), (1.4) on $[t_0, T]$. Since t_0 is fixed but otherwise arbitrary, we can always find a solution on $(-\infty, T]$. Of course Q may not be bounded on $(-\infty, T]$. If we wish to solve the infinite horizon problem (1.1), (1.2) with $T = +\infty$, then it turns out that we need a bounded solution of (1.3) on $[t_0, \infty)$. Since t_0 can vary, we require a bounded solution on $(-\infty, \infty)$. If our system is defined only on a semi-infinite interval $[T_0, \infty)$ (thus, $t_0 \geqq T_0$), then we need a bounded solution of (1.3) on $[T_0, \infty)$ (a semi-infinite interval in the *positive* direction). If all matrices are periodic with a common period θ and if (A, B) is stabilizable and (A, M) detectable, the existence of a θ-periodic solution to (1.3) is known [31], [34]. This result remains true also in infinite dimensions [12], [14]. But the existence problem for general bounded continuous matrices seems to be new.

In this paper we consider (1.3) in infinite dimensions. We assume that $A(t)$ generates an evolution operator in a Hilbert space and that other operators are bounded and continuous. We give a necessary and sufficient condition for the existence of a bounded solution to (1.3). We have uniqueness if (A, M) is detectable. If these two hypotheses are fulfilled, there exists a unique bounded solution Q_∞. We show that the optimal control for (1.1), (1.2) with $T = \infty$ is given by the usual feedback control involving Q_∞. We introduce two quadratic problems of different kind. If (1.1) is replaced by

$$y' = A(t)y + B(t)u + f(t), \tag{1.5}$$

$$y(t_0) = y_0, \tag{1.6}$$

* Received by the editors August 14, 1987; accepted for publication (in revised form) March 24, 1989.
† Scuola Normale Superiore, 56100 Pisa, Italy.
‡ Faculty of Engineering, Shizuoka University, Hamamatsu 432, Japan. This work was done while the author was a visiting professor at the Scuola Normale Superiore, 56100 Pisa, Italy.

then a more natural cost functional is

$$J_1(u) = \overline{\lim_{T\to\infty}} \frac{1}{T} \int_{t_0}^{t_0+T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt. \tag{1.7}$$

With (1.5) we also associate

$$J_2(u) = \overline{\lim_{T\to\infty}} \frac{1}{2T} \int_{-T}^{T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt. \tag{1.8}$$

We will show that Q_∞ also characterizes optimal control of these problems. This is a generalization of the average cost criterion (usually for time-invariant systems) to time-varying systems.

In § 2 we give basic assumptions on our system (1.1), (1.2).

In § 3 we establish the existence of a bounded solution to the Riccati equation (1.3). We then characterize optimal control using Q_∞. We will show that the optimal closed-loop system for (1.1) is exponentially asymptotically stable. In § 4 we consider the stochastic case and obtain similar results. We also consider the partially observable case and show that the separation principle holds [20], [40].

An important special class of time-varying systems is that of periodic systems. See [7], [8], [22], [35], and [37] for various examples of periodic systems and their optimization problems. We have studied [17] the quadratic problem (1.5)–(1.7) and its stochastic version for periodic systems. We may allow for almost periodic inputs such as in [16].

This paper is an extension of the last two papers. Hence if we assume θ-periodicity of our system, we recover earlier results in [16] and [17].

2. Preliminaries. Let Z be a Hilbert space ($\langle\,,\rangle$ inner product, $|\ |$ norm). We will denote by $\mathscr{L}(Z)$ the Banach space of all linear bounded operators $S: Z \to Z$ endowed with the norm $|S| = \sup\{|Sz|: z \in Z, |z| \leqq 1\}$. If $S \in \mathscr{L}(Z)$ then S^* will represent its adjoint operator. S is called nonnegative ($S \geqq 0$) if S is self-adjoint and $\langle Sz, z\rangle \geqq 0$, for all $z \in Z$. We set $\mathscr{L}^+(Z) = \{S \in \mathscr{L}(Z): S \geqq 0\}$. If $L: D(L) \subset Z \to Z$ is a linear operator we denote by $\sigma(L)$ (respectively, $\rho(L)$) the spectrum (respectively, the resolvent set) of L and by $R(\lambda, L)$, $\lambda \in \rho(L)$ the resolvent operator of L.

For each interval J in $\mathbb{R}^1$ we denote by $C_S(J; \mathscr{L}(Z))$ the set of all mappings $S(t): J \to \mathscr{L}(Z)$ that are strongly continuous, that is, $S(t)z$ is continuous on J for any $z \in Z$. If J is closed and bounded, then due to the Uniform Boundedness Theorem, $C_S(J; \mathscr{L}(Z))$ is a Banach space with respect to the norm:

$$\|S\| = \sup\{|S(t);\ t \in J\}.$$

We set $C_S(J; \mathscr{L}^+(Z)) = \{S \in C_S(J; \mathscr{L}(Z));\ S(t) \geqq 0,\ t \in J\}$. If X is another Hilbert space, we denote by $\mathscr{L}(X, Z)$ the set of all bounded linear operators from X into Z and by $C_S(J; \mathscr{L}(X, Z))$ the set of all strongly continuous mappings from J into $\mathscr{L}(X, Z)$.

Let Y be a Hilbert space. We consider the initial value problem

$$y' = A(t)y + f(t), \quad y(s) = y_0, \quad t \geqq s \tag{2.1}$$

where $y_0 \in Y$ and $f \in L^2_{\text{loc}}(0, \infty; Y)$, the set of locally square integrable functions. We assume the following on $A(t)$:

(H1) (i) For any $t \in \mathbb{R}^1$, $A(t)$ is a linear operator in Y with a constant domain D dense in Y. There exist numbers $\bar{M} > 0$, $\eta \in (\pi/2, \pi)$, $\delta \in \mathbb{R}^1$ such that

$$S_{\delta,\eta} = \{\lambda \in \mathbb{C}: |\arg(\lambda - \delta)| < \eta\} \subset \rho(A(t)) \quad \forall t \in \mathbb{R}^1$$

and the resolvent operator satisfies

$$|R(\lambda, A(t))| \leqq \bar{M}/|\lambda - \delta| \quad \forall \lambda \in S_{\delta,\eta}.$$

(ii) There exist numbers $\alpha \in (0, 1)$ and $\bar{N}$ such that

$$|A(t)x - A(s)x| \leqq \bar{N}|t-s|^{\alpha}|A(0)x| \quad \forall x \in D.$$

Remark 2.1. The hypothesis (H1) has been introduced by Tanabe [36] to study the abstract equation $y' = A(t)y$ where $A(t)$'s are infinitesimal generators of analytic semigroups with constant domains (parabolic equations). In fact, in the sequel we will only need the existence of an evolution operator $U(t, s)$ relative to $A(t)$. Thus our results can be easily arranged to cover hyperbolic equations as well as parabolic equations with nonconstant domains $D(A(t))$. If A is constant, (H1) is interpreted as A being the infinitesimal generator of a C_0-semigroup. Then functional differential equations can be covered [10].

The following result is proved in [36].

PROPOSITION 2.1. *Assume* (H1). *Then there exists a family of operators* $U(t, s) \in \mathscr{L}(Y)$, $t \geqq s$ *such that*

(i) $U(s, s) = I$, $s \in \mathbb{R}^1$,
(ii) $U(\cdot, \cdot)x$ *is continuous for any* $x \in Y$,
(iii) *For* $t > s$, $U(t, s)(Y) \subset D$ *and* $U(t, s)$ *is differentiable in* t *with*

$$\frac{\partial U(t, s)}{\partial t} = A(t)U(t, s).$$

$U(t, s)$ is called the *evolution operator* relative to $A(t)$. It is called (*exponentially*) *stable* if there exist positive numbers $\tilde{M}$, ω such that $|U(t, s)| \leqq \tilde{M} e^{-\omega(t-s)}$ for any $t \geqq s$.

We denote by $A_n(t) = n^2 R(n, A(t)) - nI$ the Yosida approximations of $A(t)$ and by $U_n(t, s)$ the evolution operator relative to $A_n(t)$ (that clearly exists since $A_n(t)$'s are bounded). By using the results in [36] it is easy to prove that

$$\lim_{n\to\infty} U_n(t, s)x = U(t, s)x \quad \forall t \geqq s, \quad \forall x \in Y \tag{2.2}$$

uniformly on the bounded sets of $\mathbb{R}^2$.

We define the *mild solution* of (2.1) by

$$y(t) = U(t, s)y_0 + \int_s^t U(t, r)f(r)\, dr. \tag{2.3}$$

It is continuous on $[s, \infty)$. Let y_n be the classical solution to the problem:

$$y_n' = A_n(t)y_n + f(t), \qquad y_n(s) = y_0. \tag{2.4}$$

Then, by (2.2) $y_n(t) \to y(t)$ uniformly on any bounded subset of $[s, \infty)$.

Assume that $A(t)$ is stable. Then for each $f \in L^{\infty}([s, \infty); Y)$ (the set of bounded measurable functions in Y) $y(t)$ defined by (2.3) is bounded. We now consider for each $f \in L^{\infty}(\mathbb{R}^1; Y)$

$$y' = A(t)y + f(t) \tag{2.4'}$$

on $[-\infty, \infty)$. We say that $y(t)$ is a mild solution on $(-\infty, \infty)$ if it satisfies the integral equation

$$y(t) = U(t, s)y(s) + \int_s^t U(t, r)f(r)\, dr \tag{2.5}$$

for any $t \geqq s$. A mild solution $y(t)$ is called *bounded* if it is bounded on $(-\infty, \infty)$. If $A(t)$ is stable, then there exists a unique bounded mild solution of (2.4'). In fact it is given by

$$y(t) = \int_{-\infty}^{t} U(t, r) f(r)\, dr. \tag{2.6}$$

If A, f are θ-periodic, then y is also θ-periodic. If A is θ-periodic and f is almost periodic, then y is almost periodic [16].

3. Optimal quadratic control in the deterministic case.

3.1. Bounded solutions of a Riccati equation. We consider the usual quadratic control problem.

$$y' = A(t)y + B(t)u, \qquad y(t_0) = y_0, \tag{3.1}$$

$$J_0(u) = \int_{t_0}^{\infty} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt \tag{3.2}$$

where $A(t)$ satisfies the condition (H1) and

(H2) (i) $B \in C_S(\mathbb{R}^1, \mathscr{L}(U, Y)) \cap L^\infty(\mathbb{R}^1, \mathscr{L}(U, Y))$, $M \in C_S(\mathbb{R}^1; \mathscr{L}(Y))$, $N \in C_S(\mathbb{R}^1; \mathscr{L}^+(U))$ and there exists an $\varepsilon > 0$ such that $N(t) \geqq \varepsilon$ for any $t \in \mathbb{R}^1$.

(ii) $\sup_{t\in\mathbb{R}} [|M(t)| + |N(t)|] < \infty$.

We wish to minimize $J(u)$ over the set of admissible controls

$$\mathscr{U}_{\text{ad}}^0 = \{u \in L^2([t_0, \infty); U)\text{: the corresponding mild solution } y(t) \to 0 \text{ as } t \to \infty\}. \tag{3.3}$$

To solve this problem the following Riccati equation is useful:

$$Q'(t) + A^*(t)Q(t) + Q(t)A(t) + M^*(t)M(t) - Q(t)B(t)N^{-1}(t)B^*(t)Q(t) = 0. \tag{3.4}$$

We say that Q is a mild solution of (3.4) on the interval $J \subset \mathbb{R}^1$ if $Q \in C_S(J, \mathscr{L}^+(Y))$ and if it satisfies the integral equation

$$\begin{aligned} Q(t)x &= U^*(s, t)Q(s)U(s, t)x \\ &\quad + \int_t^s U^*(r, t)[M^*(r)M(r) - Q(r)B(r)N^{-1}(r)B^*(r)Q(r)]U(r, t)x\, dr \end{aligned} \tag{3.5}$$

for any $x \in Y$ and $t \leqq s$, $t, s \in J$. If $\sup_{t\in\mathbb{R}^1} |Q(t)| < \infty$, we say that Q is a *bounded solution* of (3.4). Even if Q is a solution of the integral equation (3.5) we cannot in general prove that $Q(t)x$ is differentiable for $x \in Y$. Thus Q is not a classical solution to (3.4). Therefore it is useful to introduce approximating systems

$$Q_n' + A_n^* Q_n + Q_n A_n + M^*M - Q_n B N^{-1} B^* Q_n = 0, \tag{3.6}$$

which have classical solutions. The following result is proved in [4].

PROPOSITION 3.1. *Assume* (H1) *and* (H2). *Let* $T \in \mathbb{R}^1$ *and* $Q_0 \in \mathscr{L}^+(Y)$. *Then there exists a unique mild solution* Q *of* (3.4) *on* $(-\infty, T]$ *such that* $Q(T) = Q_0$. *Moreover, there exists a unique classical solution* $Q_n \in C_S((-\infty, T]; \mathscr{L}^+(Y))$ *to* (3.6) *with* $Q_n(T) = Q_0$ *and* $Q_n(t)x \to Q(t)x$ *as* $n \to \infty$ *for any* $x \in Y$ *uniformly on any bounded subset of* $(-\infty, T]$.

In the sequel we set $Q(t) = \Lambda(t; T, Q_0)$, $Q_n(t) = \Lambda_n(t; T, Q_0)$. The following monotonicity property of $\Lambda(\Lambda_n)$ is well known:

$$\begin{aligned} &\Lambda(t; T, Q_0) \leqq \Lambda(t; T, Q_1), \\ &\Lambda_n(t, T, Q_0) \leqq \Lambda_n(t, T, Q_1) \quad \text{if } Q_0 \leqq Q_1. \end{aligned} \tag{3.7}$$

Now we will establish a bounded solution to (3.4). Let $C_b(\mathbb{R}^1; Z)$ be the space of all bounded continuous functions from $\mathbb{R}^1$ to Z. The fundamental hypothesis for the existence of a bounded solution is the following:

(H3) For any $t_0 \in \mathbb{R}$ and $y_0 \in Y$ there exist $u \in C_b(\mathbb{R}^1; U)$ and $C_0 > 0$ such that

$$\int_{t_0}^{\infty} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt \leqq C_0 |y_0|^2$$

where y is the mild solution to (3.1).

Hypothesis (H3) is slightly weaker than the existence of an admissible control. It is satisfied, as we will see below, if (A, B) is *stabilizable*, that is, if there exists $K \in C_S(\mathbb{R}^1; \mathscr{L}(Y, U))$ bounded such that the evolution operator relative to $A - BK$ is stable (such an evolution operator does exist since BK is bounded [10]).

Assume that (A, B) is stabilizable so that for some K

$$|U_K(t, s)| \leqq M_0\, e^{-\omega(t-s)}, \qquad t \geqq s \quad \text{for some } M_0 \geqq 1 \text{ and } \omega > 0 \tag{3.8}$$

where U_K is the evolution operator relative to $A - BK$. Now we will show that the hypothesis (H3) is fulfilled. Set

$$y(t) = U_K(t, t_0)y_0, \qquad u(t) = -K(t)U_K(t, t_0)y_0.$$

Then

$$\int_{t_0}^{\infty} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt \leqq \frac{M_0^2}{2\omega}(\|M\|^2 + \|N\|\,\|K\|^2)$$

where $\|\cdot\| = \sup_{t \in \mathbb{R}^1} |\cdot|$.

The main result for our Riccati equation is the following theorem.

THEOREM 3.1. *Assume* (H1) *and* (H2). *Then a nonnegative bounded solution to* (3.4) *exists if and only if* (H3) *holds.*

Proof. If. Assume (H3). For any $\alpha \in \mathbb{R}^1$ set $Q_\alpha = \Lambda(\cdot; \alpha, 0)$. By (3.7) we have

$$Q_\alpha(t) \leqq Q_\beta(t) \quad \text{if } t \in (-\infty, \alpha] \text{ and } \alpha \leqq \beta. \tag{3.9}$$

Thus $\{Q_\alpha\}$ is increasing in α. We will now show that $\|Q_\alpha(\cdot)\|$ is bounded. Let $Q_{\alpha,n} = \Lambda_n(\cdot; \alpha, 0)$ and let y_n be the classical solution to the initial value problem:

$$y_n' = A_n(t)y_n + B(t)u, \qquad y_n(t_0) = y_0 \tag{3.10}$$

where u is the control function given in (H3). We then have

$$\begin{aligned} \frac{d}{dt}\langle Q_{\alpha,n}(t)y_n(t), y_n(t)\rangle &= \|N^{1/2}(u + N^{-1}B^*Q_{\alpha,n}y_n)|^2 - |M(t)y_n(t)|^2 \\ &\quad - \langle N(t)u(t), u(t)\rangle. \end{aligned} \tag{3.11}$$

Integrating this from t_0 to α and letting $n \to \infty$, we arrive at

$$\begin{aligned} &\int_{t_0}^{\alpha} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt \\ &\quad = \langle Q_{\alpha,n}(t_0)y_0, y_0\rangle + \int_{t_0}^{\alpha} |N^{1/2}(u + N^{-1}B^*Q_{\alpha,n}y_n)|^2\, dt, \end{aligned} \tag{3.12}$$

which yields

$$\langle Q_\alpha(t_0)y_0, y_0\rangle \leqq C_0|y_0|^2 \quad \text{for any } y_0 \in Y. \tag{3.13}$$

By a classical argument we can show that there exists $Q_\infty(t)$ such that

$$\lim_{\alpha\to\infty} Q_\alpha(t_0)y_0 = Q_\infty(t_0)y_0 \quad \text{for any } t_0\in\mathbb{R}^1 \text{ and } y_0\in Y. \tag{3.13'}$$

To prove that Q_∞ is a mild solution of (3.4) on $(-\infty,\infty)$ it suffices to let $\alpha\to\infty$ in the equality:

$$\begin{aligned} Q_\alpha(t)x = {} & U^*(s,t)Q_\alpha(s)U(s,t) \\ & + \int_t^s U^*(r,t)[M^*(r)M(r) \\ & - Q_\alpha(r)BN^{-1}BQ_\alpha(r)]U(r,t)\,dr \end{aligned} \tag{3.14}$$

for $t\leqq s<\alpha$.

Only if. Let Q be a bounded solution to (3.4). For a fixed but otherwise arbitrary $T\in\mathbb{R}^1$, let $Q_n=\Lambda_n(\cdot, T; Q(T))$. Set $K=BN^{-1}B^*$, $L=A-KQ$, $L_n=A_n-KQ_n$ and let U_L and U_{L_n} be evolution operators relative to L and L_n, respectively. Then

$$Q_n'+L_n^*Q_n+Q_nL_n+M^*M+Q_nKQ_n=0, \qquad Q_n(T)=Q(T). \tag{3.15}$$

Hence, for any $t_0\leqq t\leqq T$, we have

$$\begin{aligned} & \frac{d}{dt}\langle Q_n(t)U_{L_n}(t,t_0)y_0,\, U_{L_n}(t,t_0)y_0\rangle \\ & \quad = -|M(t)U_{L_n}(t,t_0)y_0|^2-|\sqrt{K(t)}Q_n(t)U_{L_n}(t,t_0)y_0|^2. \end{aligned} \tag{3.16}$$

Integrating this from t_0 to t_1 and letting $n\to\infty$, we obtain

$$\begin{aligned} & \int_{t_2}^{t_1}[|M(t)U_L(t,t_0)y_0|^2+|\sqrt{K(t)}Q(t)U_L(t,t_0)y_0|^2]\,dt \\ & \quad +\langle Q(t_1)U_L(t_1,t_0)y_0, U_L(t_1,t_0)y_0\rangle = \langle Q(t_0)y_0, y_0\rangle. \end{aligned} \tag{3.17}$$

Now set

$$y(t)=U_L(t,t_0)y_0, \qquad u(t)=-N^{-1}(t)B^*(t)Q(t)U_L(t,t_0)y_0.$$

Then $u\in C_b(\mathbb{R}^1; U)$ and y is a mild solution of (3.1). Moreover,

$$\int_{t_0}^{t_1}[|M(t)y|^2+\langle N(t)u,u\rangle]\,dt \leqq \sup_{t_0\in\mathbb{R}}|Q(t_0)||y_0|^2, \qquad t_0\leqq t_1\leqq T.$$

Since t_1 is arbitrary, we have shown (H3). □

If (H1)-(H3) hold, we will denote by Q_∞ the bounded solution of (3.4) defined by (3.13). We remark that Q_∞ is minimal among all solutions $Q\geqq 0$ of (3.4) on $\mathbb{R}^1$, that is,

$$Q(t)\geqq Q_\infty(t), \qquad t\in\mathbb{R}^1. \tag{3.18}$$

In fact if Q is a solution of (3.4) on $\mathbb{R}^1$, we have

$$Q(\alpha)\geqq Q_\alpha(\alpha)=0$$

so that

$$Q(t)\geqq Q_\alpha(t) \quad \text{for any } t\in(-\infty,\alpha].$$

Letting $\alpha\to\infty$ we obtain (3.18). We will call Q_∞ the *minimal solution* of (3.4) on $\mathbb{R}^1$.

Next we will examine the stability property of a bounded solution Q of (3.4). Set $L = A - KQ$, $K = BN^{-1}B^*$ and for a fixed $T \in \mathbb{R}^1$ let $Q_1(\cdot) = \Lambda(\cdot; T, S)$, where $S \in \mathscr{L}^+(Y)$. Then we can easily check that $Z = Q_1 - Q$ is a mild solution of the equation

$$Z' + L^*Z + ZL - ZKZ = 0, \qquad Z(T) = S - Q(T). \tag{3.19}$$

If U_L is stable, the usual linearization arguments show that Q is uniformly asymptotically stable as $t \to -\infty$ [23]. But as we will see below we can show that Q is attractive from above, and hence it is maximal among all bounded nonnegative solutions of (3.4). This will imply, in particular, the uniqueness of a nonnegative bounded solution for which L is stable. We say that a bounded solution Q of (3.4) is *stable* if $A - KQ$ is stable.

PROPOSITION 3.2. *Assume* (H1) *and* (H2) *and let Q be a stable bounded nonnegative solution to* (3.4). *Let $T \in \mathbb{R}^1$, $S = \mathscr{L}^+(Y)$ be arbitrary and set $Q_1(\cdot) = \Lambda(\cdot; T, S)$. If $S \geqq Q(T)$, then*

$$\lim_{t \to -\infty} (Q_1(t)x - Q(t)x) = 0 \quad \textit{for any } x \in Y.$$

$Q_1(\cdot)$ with arbitrary $S \geqq 0$ is bounded on $(-\infty, T]$.

Moreover, if Q_2 is a bounded solution, then $Q_2(t) \leqq Q(t)$, $t \in \mathbb{R}$.

Proof. Let $Z = Q_1 - Q$, $Q_n = \Lambda_n(\cdot; T, Q(T))$, $Q_{1n} = \Lambda_n(\cdot; T, S)$, and define $Z_n = Q_{1n} - Q_n$, $L_n = A_n - KQ_n$. Then

$$Z_n' + L_n^* Z_n + Z_n L_n - Z_n K Z_n = 0, \tag{3.20}$$

from which follows

$$\frac{d}{dt}\langle Z_n(t) U_{L_n}(t, t_0)y_0, U_{L_n}(t, t_0)y_0\rangle = |\sqrt{K(t)}\, Z_n(t) U_{L_n}(t, t_0)y_0|^2. \tag{3.21}$$

Integrating from t_0 to t and letting $n \to \infty$, we obtain

$$\langle Z(t) U_L(t, t_0)y_0, U_L(t, t_0)y_0\rangle \geqq \langle Z(t_0)y_0, y_0\rangle, \qquad t_0 \leqq t. \tag{3.22}$$

Now, if $S \geqq Q(T)$ then $Z(t_0) \geqq 0$ for any $t_0 \leqq T$. Letting $t_0 \to -\infty$ in (3.22), we obtain $\langle Z(t_0)y_0, y_0\rangle \to 0$ as $t_0 \to -\infty$. Hence $\lim_{t \to -\infty} Z(t)x = 0$ for any $x \in Y$. Assume now that Q_2 is another nonnegative bounded solution of (3.4). Then replacing Q_1 by Q_2 in (3.22) and letting $t \to \infty$ we find $\langle Z(t_0)y_0, y_0\rangle \leqq 0$ so that $Q_2(t) \leqq Q(t)$. □

Now we give a sufficient condition for a bounded nonnegative solution of (3.4) being stable:

(H4) There exists a $K_1 \in C_S(\mathbb{R}^1; \mathscr{L}(Y))$ bounded such that $A - K_1 M$ is stable.

If (H4) holds, we say that (A, M) is *detectable.*

PROPOSITION 3.3. *Assume* (H1), (H2), *and* (H4). *Then any bounded nonnegative solution of* (3.4) *is stable. Thus the Riccati equation* (3.4) *has at most one bounded nonnegative solution.*

Proof. Let $t_0 \in \mathbb{R}$ be fixed and let $y_0 \in Y$. Then by (3.17) we have

$$M(t) U_L(t, t_0)y_0, \qquad \sqrt{K(t)}\, Q(t) U(t, t_0)y_0 \in L^2(t_0, \infty; Y). \tag{3.23}$$

Let $S = A - K_1 M$; then $L = S + K_1 M - KQ$ so that

$$U_L(t, t_0)y_0 = U_S(t, t_0)y_0 - \int_{t_0}^{t} U_S(t, r)(K_1 M - KQ) U_L(r, t_0)y_0\, dr. \tag{3.24}$$

Since U_S is stable, it follows from (3.23), (3.24) that $U_L(t, t_0)y_0 \in L^2(t_0, \infty; Y)$. By Datko [19] U_L is stable.

The uniqueness follows from the fact that a stable solution is maximal. □

In practice it may be more natural to assume that the system (3.1) is defined only on $[T_0, \infty)$ so that $T_0 \geqq T_0 < \infty$. In this case we restrict the hypothesis (H1)-(H4) on $[T_0, \infty)$. We need to modify the definitions.

For example, we say that (A, B) is stabilizable if there exists $K \in C_S([T_0, \infty); L(Y, U))$ bounded such that $|U_{A-BK}(t, s)| \leqq M_0 e^{-\omega(t-s)}$, $t \geqq s \geqq T_0$ for some $M_0 \geqq 1$ and $\omega > 0$. Now all the results restricted on $[T_0, \infty)$ are true. In fact, we have the following corollaries.

COROLLARY 3.1. *Assume* (H1) *and* (H2) *on* $[T_0, \infty)$. *Then a nonnegative bounded solution of* (3.4) *on* $[T_0, \infty)$ *exists if and only if* (H3) *holds on* $[T_0, \infty)$.

COROLLARY 3.2. *Assume* (H1) *and* (H2) *on* $[T_0, \infty)$. *Let* Q *be a stable nonnegative bounded solution of* (3.4) *on* $[T_0, \infty)$. *Then for any bounded solution* $Q_2 \geqq 0$ *of* (3.4), $Q_2(t) \leqq Q(t)$, $t \in [T_0, \infty)$, *that is* Q *is maximal.*

COROLLARY 3.3. *Assume* (H1), (H2), *and* (H4) *on* $[T_0, \infty)$. *Then any nonnegative bounded solution of* (3.4) *is stable. Thus the Riccati equation* (3.4) *has at most one nonnegative bounded solution on* $[T_0, \infty)$.

Finally, we consider two special cases of (3.4): the periodic case and the time invariant case.

In the former we assume (H5).

(H5) There exists a number $\theta > 0$ such that $A(t+\theta) = A(t)$, $B(t+\theta) = B(t)$, $M(t+\theta) = M(t)$, and $N(t+\theta) = N(t)$ for all $t \in \mathbb{R}^1$.

In this case we say that these operators are θ-periodic. We say also that the system (3.1) is θ-periodic. If Q is a bounded solution to (3.4), then $Q_\theta(t)$ defined by

$$Q_\theta(t) = Q(t-\theta), \qquad t \in \mathbb{R}^1$$

is also a bounded solution. Thus if (3.4) has a unique nonnegative bounded solution Q (for example, if (A, M) is detectable) we have $Q(t) = Q(t-\theta)$. In fact, we have Proposition 3.4.

PROPOSITION 3.4. *Assume* (H1), (H3), *and* (H5). *Then the minimal solution* Q_∞ *of* (3.4) *is* θ*-periodic. If, further,* (H4) *holds, then* Q_∞ *is the unique nonnegative* θ*-periodic solution to* (3.4) *and it is uniformly asymptotically stable.*

Proof. Let n be an integer and set

$$V(t) = Q_{n\theta}(t-\theta), \qquad t \in (-\infty, (n+1)\theta] \tag{3.25}$$

where $Q_{n\theta} = \Lambda(\cdot; n\theta, 0)$. Since the coefficients of (3.4) are θ-periodic, V is also a solution of (3.4) on $(-\infty, (n+1)\theta]$. Moreover, $V((n+1)\theta) = Q_{n\theta}(n\theta) = 0$ so that

$$V(t) = Q_{(n+1)\theta}(t) = Q_{n\theta}(t-\theta). \tag{3.26}$$

Now, letting $n \to \infty$ we obtain $Q_\infty(t) = Q_\infty(t-\theta)$. Thus Q_∞ is θ-periodic. Other assertions follow from Proposition 3.2. □

Next we show the global orbital attractiveness of Q_∞.

PROPOSITION 3.5. *Assume* (H1)-(H5). *Let* $S_0 \in \mathscr{L}^+(Y)$ *and set* $Q = \Lambda(\cdot; 0, S_0)$. *Then*

$$\lim_{n\to\infty} Q(t-n\theta)x = Q_\infty(t)x \quad \forall t \in (-\infty, 0]. \tag{3.27}$$

Proof. Let m be an integer such that

$$S_0 \leqq mI \quad \text{and} \quad Q_\infty(0) \leqq mI.$$

Let $V(\cdot) = \Lambda(\cdot; 0, mI)$. Then

$$Q(t) \geqq Q_0(t) \quad \forall t \leqq 0,$$

$$Q(t - n\theta) \geqq Q_0(t - n\theta) = Q_{n\theta}(t) \quad \forall t \leqq n\theta.$$

This implies

$$Q_{n\theta}(t) \leqq Q(t - n\theta) \leqq V(t - n\theta).$$

But $Q_{n\theta}(t)x \to Q_\infty(t)x$ as $n \to \infty$ and $V(t - n\theta)x \to Q_\infty(t)x$ by Proposition 3.2. Thus (3.27) follows. □

Remark 3.1. In Da Prato [12] the existence of a periodic solution to (3.4) is shown under (H1), (H2), (H5), and stabilizability of (A, B). Hence, Proposition 3.4 gives a weaker condition. Proposition 3.5 is also an improvement of Lemma 3.1 [17]. See [31] and [34] for finite-dimensional results.

Now consider the time-invariant case: A, B, M, and N are independent of t. Then (H1) can be replaced by the hypothesis that A is the infinitesimal generator of a C_0-semigroup e^{tA}. Hypothesis (H2) simply implies $B \in \mathscr{L}(U, Y)$, $M \in \mathscr{L}(Y)$, and $N, N^{-1} \in \mathscr{L}^+(U)$. Hypothesis (H4) is the usual detectability condition [39], [41]. Hence we recover the results of Zabczyk [41].

PROPOSITION 3.6. *Suppose that A is the infinitesimal generator of a C_0-semigroup e^{tA} and that $B \in \mathscr{L}(U, Y)$, $M \in \mathscr{L}(Y)$, and $N, N^{-1} \in \mathscr{L}^+(U)$. Suppose (H3) holds. Then $Q_\infty(t) = Q_\infty$ is independent of t and is the minimal solution of the algebraic Riccati equation*

$$A^*Q + QA + M^*M - QBN^{-1}B^*Q = 0. \tag{3.28}$$

If, further, (A, M) is detectable, then Q_∞ is the unique nonnegative solution to (3.20) *in $\mathscr{L}^+(Y)$. Moreover, for each $Q = \Lambda(\cdot, \cdot, S_0)$ with $S_0 \in \mathscr{L}^+(Y)$*

$$\overline{\lim_{t \to -\infty}}\, Q(t)x = Q_\infty x \quad \forall x \in Y.$$

3.2. Quadratic control on the infinite horizon. Let $-\infty \leqq T_0 < \infty$ and let $t_0 \in [T_0, \infty)$ be arbitrary. If $T_0 = -\infty$, we mean by $[T_0, \infty)$, the whole real line $(-\infty, \infty)$. Now consider our control problem

$$y' = A(t)y + B(t)u, \qquad y(t_0) = y_0, \tag{3.1}$$

$$J_0(u) = \int_{t_0}^{\infty} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt. \tag{3.2}$$

We assume (H1)-(H4) on $[T_0, \infty)$ and wish to minimize $J_0(u)$ over

$$\mathscr{U}_{\text{ad}}^0 = \{u \in L^2(t_0, \infty; U)\text{: the corresponding mild solution } y(t) \to 0 \text{ as } t \to \infty\}. \tag{3.3}$$

In view of (H3), (H4), this problem is nontrivial. Let Q_∞ be the unique stable nonnegative bounded solution of (3.4) on $[T_0, \infty)$. The main tool we will use is the following identity.

LEMMA 3.1. *Assume* (H1)-(H3) *on* $[T_0, \infty)$. *Let* $u \in L^2([t_1, t_2]; U)$, $t_0 \leqq t_1 < t_2$ *and let y be the mild solution of* (3.1) *on* $[t_1, t_2]$. *Then*

$$\begin{aligned} &\int_{t_1}^{t_2} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt + \langle Q_\infty(t_2)y(t_2), y(t_2)\rangle \\ &\quad = \int_{t_1}^{t_2} |N^{1/2}(u + N^{-1}B^*Q_\infty y)|^2\, dt + \langle Q_\infty(t_1)y(t_1), y(t_1)\rangle. \end{aligned} \tag{3.29}$$

Proof. Let $Q_n = \Lambda_n(\cdot\,; t_2, Q_\infty(t_2))$ and let y_n be the solution of the initial value problem

$$y_n' = A_n(t)y_n + B(t)u, \qquad y_n(t_1) = y(t_1).$$

Then we have

$$\frac{d}{dt}\langle Q_n(t)y_n(t), y_n(t)\rangle = |N^{1/2}(u + N^{-1}B^*Q_n y_n)|^2 - |M(t)y_n|^2 - \langle N(t)u, u\rangle.$$

Integrating this from t_1 and t_2 and passing to the limit $n \to \infty$, we obtain (3.29).

Now it is easy to solve our control problem.

THEOREM 3.2. *Assume* (H1)-(H4) *on* $[T_0, \infty)$. *Then the optimal control is given by the feedback law*

$$\bar{u} = -N^{-1}B^*Q_\infty \bar{y} \tag{3.30}$$

and the optimal cost by

$$J_0(\bar{u}) = \langle Q_\infty(t_0)y_0, y_0\rangle. \tag{3.31}$$

The optimal closed loop system is stable. If, further, (H5) *holds, then* Q_∞ *is* θ*-periodic. If all operators in* (3.1), (3.2) *are time invariant, then* Q_∞ *is constant.*

Proof. We set $t_1 = t_0$ and pass to the limit $t_2 \to \infty$. Since $y(t_2) \to 0$ as $t_2 \to \infty$ we obtain

$$J_0(u) = \int_{t_0}^{\infty} |N^{1/2}(u + N^{-1}B^*Q_\infty y)|^2\, dt + \langle Q_\infty(t_0)y_0, y_0\rangle.$$

Thus the conclusion follows immediately. □

3.3. The optimal control problem with average cost. Assume (H1), (H2) on $[T_0, \infty)$. Here we are concerned with a more general system

$$y' = A(t)y + B(t)u + f(t), \qquad y(t_0) = y_0 \tag{3.32}$$

where $t_0 \in [T_0, \infty)$ is fixed but otherwise arbitrary and $f \in C_b([T_0, \infty), Y)$. In this case we cannot expect that the cost $J_0(u)$ is finite. Instead we take a more reasonable cost

$$J_1(u) = \overline{\lim_{T\to\infty}} \frac{1}{T} \int_{t_0}^{t_0+T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt \tag{3.33}$$

and we wish to minimize it over

$$\mathcal{U}'_{ad} = \left\{ u: \overline{\lim_{T\to\infty}} \frac{1}{T} \int_{t_0}^{t_0+T} |u(t)|^2\, dt < \infty \text{ and the mild solution } y(t) \text{ of (3.32) is bounded on } [t_0, \infty) \right\}. \tag{3.34}$$

We further assume (H3) and let Q_∞ be the minimal bounded solution of (3.4) on $[T_0, \infty)$.

Let $L = A - KQ_\infty$, $K = BN^{-1}B^*$ and consider the following equation:

$$r' + L^*r + Q_\infty f = 0. \tag{3.35}$$

Then we have a similar result to Lemma 3.1.

LEMMA 3.2. *Assume* (H1)-(H3) *on* $[T_0, \infty)$. *Let* $u \in L^2([t_1, t_2]; U)$, $t_0 \leqq t_1 < t_2$, *and let* y *and* r *be mild solutions of* (3.32), (3.35) *on* $|t_1, t_2|$, *respectively. Then*

$$\begin{aligned}\int_{t_1}^{t_2} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt + \langle Q_\infty(t_2)y(t_2), y(t_2)\rangle + 2\langle r(t_2), y(t_2)\rangle \\ = \int_{t_1}^{t_2} |N^{1/2}[u + N^{-1}B^*(Q_\infty y + r)]|^2\, dt \\ + \int_{t_1}^{t_2} [2\langle r, f\rangle - |N^{-1/2}B^* r|^2]\, dt \\ + \langle Q_\infty(t_1)y(t_1), y(t_1)\rangle + 2\langle r(t_1), y(t_1)\rangle.\end{aligned} \tag{3.36}$$

Proof. We take

$$\begin{aligned} &Q_n(\cdot) = \Lambda_n(\cdot\,; t_2, Q_\infty(t_2)), \\ &y_n' = A_n y_n + Bu + f, \qquad y_n(t_1) = y(t_1), \\ &r_n' + (A_n - KQ_n)^* r_n + Q_n f = 0, \qquad r_n(t_2) = r(t_2)\end{aligned}$$

and show

$$\begin{aligned}\frac{d}{dt}[\langle Q_n y_n, y_n\rangle + 2\langle r_n, y_n\rangle] = |N^{1/2}[u + N^{-1}B^*(q_n y_n + r_n)]|^2 \\ - |My_n|^2 - \langle Nu, u\rangle + 2\langle r_n, f_n\rangle - |N^{-1/2}B^* r_n|^2.\end{aligned}$$

Integrating this from t_1 to t_2 and passing to the limit $n \to \infty$, we obtain (3.36).

It we further assume (H4) on $[T_0, \infty)$, then $\mathcal{U}'_{\mathrm{ad}}$ is not empty. To see this, note that there exists a unique bounded solution to (3.35) given by

$$r(t) = \int_t^\infty U_L^*(s, t) Q_\infty(s) f(s)\, ds, \tag{3.37}$$

since L is stable. Now consider the feedback control

$$\bar{u} = -N^{-1}B^*(Q_\infty \bar{y} + r). \tag{3.38}$$

Then the closed-loop system is

$$\bar{y}' = L\bar{y} + f - Kr, \qquad \bar{y}(t_0) = y_0. \tag{3.39}$$

Since L is stable, the mild solution

$$\bar{y}(t) = U_L(t, t_0)y_0 + \int_{t_0}^t U_L(t, s)[f(s) - K(s)r(s)]\, ds \tag{3.40}$$

is bounded on $[t_0, \infty)$. Hence u is admissible.

THEOREM 3.3. *Assume* (H1)-(H4) *on* $[T_0, \infty)$. *Then the optimal control is given by the feedback law* (3.38) *and the optimal cost by*

$$J_1(\bar{u}) = \overline{\lim_{T\to\infty}} \frac{1}{T} \int_{t_0}^{t_0+T} [2\langle r, f\rangle - |N^{-1/2}B^* N|^2]\, dt. \tag{3.41}$$

Proof. We take any $u \in \mathcal{U}'_{\mathrm{ad}}$ and its response y in (3.36). Then, setting $t_1 = t_0$, $t_2 = t_0 + T$ and taking limit supremum as $T \to \infty$, we obtain

$$J_1(\bar{u}) = \overline{\lim_{T\to\infty}} \frac{1}{T} \int_{t_0}^{t_0+T} \{|N^{1/2}[u + N^{-1}B^*(Qy + r)]|^2 + 2\langle r, f\rangle - |N^{-1/2}B^* r|^2\}\, dt. \tag{3.42}$$

Now the assertion follows easily. □

If we assume (H5), then we recover the periodic result in [16].

COROLLARY 3.4. *Assume* (H1)-(H4) *on* $[T_0, \infty)$.

(i) *If* (H5) *holds on* $[T_0, \infty)$ *and if* f *is* θ*-periodic, then* Q_∞ *and* r *are* θ*-periodic and*

$$J_1(\bar{u}) = \frac{1}{\theta}\int_{t_0}^{t_0+\theta} [2\langle r, f\rangle - |N^{-1/2}B^*r|^2]\,dt \quad \text{for any } t_0 \geqq T_0. \tag{3.43}$$

(ii) *If all operators in* (3.32), (3.33), *and* f *are constant, then* Q_∞ *is the unique solution of the algebraic Riccati equation* (3.28) *and*

$$J_1(\bar{u}) = 2\langle r, f\rangle - |N^{-1/2}B^*r|^2 \tag{3.44}$$

where $r = -(L^*)^{-1}Q_\infty f$.

3.4. The optimal control problem with average cost II. Here we assume (H1) and (H2). Our system is

$$y' = A(t)y + B(t)u + f(t) \tag{3.45}$$

where $f \in C_b(\mathbb{R}^1; Y)$. We wish to minimize the average cost

$$J_2(u) = \overline{\lim_{T\to\infty}} \frac{1}{2T}\int_{-T}^{T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\,dt, \tag{3.46}$$

over

$$\mathcal{U}_{\text{ad}}^2 = \left\{ u: \overline{\lim_{T\to\infty}} \frac{1}{2T}\int_{-T}^{T} |u(t)|^2\,dt < \infty \right. \tag{3.47}$$
$$\left. \text{such that there exists a bounded solution to (3.45)} \right\}.$$

If (H3) holds, then Lemma 3.2 is valid. If, further, (H4) holds, then L is stable and there exists a unique bounded solution to (3.35). Thus as in Theorem 3.3 we have Theorem 3.4.

THEOREM 3.4. *Assume* (H1)-(H4). *Then optimal control is given by the feedback law*

$$\bar{u} = -N^{-1}B^*(Q_\infty \bar{y} + r) \tag{3.48}$$

where Q_∞ *is the unique bounded nonnegative stable solution to* (3.4) *and* r *is the unique bounded solution on* $\mathbb{R}^1$ *of the equation*

$$r' + L^*r + Q_\infty f = 0 \tag{3.49}$$

given by

$$r(t) = \int_t^\infty U_L^*(s, t)Q_\infty(s)f(s)\,ds, \qquad t \in \mathbb{R}^1. \tag{3.50}$$

The optimal cost is given by

$$J_2(\bar{u}) = \overline{\lim_{T\to\infty}} \frac{1}{2T}\int_{-T}^{T} [2\langle r, f\rangle - |N^{-1/2}B^*r|^2]\,dt. \tag{3.51}$$

The optimal response $\bar{y}$ *is*

$$\bar{y}(t) = \int_{-\infty}^{t} U_L(t, s)[f(s) - K(s)r(s)]\,ds \tag{3.52}$$

and y is exponentially asymptotically stable (i.e., $y(t; t_0, y_0) - y(t) \to 0$, as $t \to \infty$ where $y(t; t_0, y_0)$ is the solution of (3.39)).

COROLLARY 3.5. *Assume* (H1)-(H4).

(i) *If* (H5) *also holds and if f is θ-periodic, then Q_∞ and r are θ-periodic and*

$$J_2(\bar{u}) = \frac{1}{\theta}\int_0^\theta [2\langle r, f\rangle - |N^{-1/2}B^*r|^2]\,dt = J_1(\bar{u}). \tag{3.53}$$

(ii) *If all operators and f are constant, then Q_∞, r are constant and given as in Corollary* 3.4. *Moreover,*

$$J_2(\bar{u}) = 2\langle r, f\rangle - |N^{-1/2}B^*r|^2 = J_1(\bar{u}). \tag{3.54}$$

Finally, we will consider another special case of Theorem 3.4. Let $AP(\mathbb{R}^1; Z)$ be the Banach space of almost periodic functions in Z [1], [16], [21]. We assume $f \in AP(\mathbb{R}^1; Y)$. We assume (H1)-(H5) so that Q_∞ is the unique nonnegative θ-periodic solution of (3.4). Then L is stable and $r(t)$, given by (3.50), is the unique almost periodic solution of (3.45). Moreover $\bar{y}$, given in (3.52), is also the unique almost periodic solution of the closed system

$$\bar{y}' = L\bar{y} + f - Kr. \tag{3.55}$$

Hence the following problem is meaningful [16]. Minimize

$$J_{\text{ap}}(u) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\,dt \tag{3.56}$$

over the set of admissible controls

$$\begin{aligned}\mathcal{U}_{\text{ap}} = \{u \in AP(\mathbb{R}^1, U)&\text{: there exists } y \in AP(\mathcal{R}^1; Y)\\ &\text{which is a mild solution of (3.45)}\}.\end{aligned} \tag{3.57}$$

Now we find the optimal almost periodic control given in [16].

COROLLARY 3.6. *Assume* (H1)-(H5) *and let $f \in AP(\mathbb{R}^1; Y)$. Then the optimal control is given by the feedback law*

$$\bar{u} = -N^{-1}B^*(Q_\infty\bar{y} + r) \tag{3.58}$$

where Q_∞ is the nonnegative θ-periodic solution to (3.4) *and r is the unique almost periodic solution to*

$$r' + L^*r + Q_\infty f = 0 \tag{3.59}$$

given by

$$r(t) = \int_t^\infty U_L^*(s, t)Q_\infty(s)f(s)\,ds. \tag{3.60}$$

The optimal cost is given by

$$J_{\text{ap}}(\bar{u}) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} [2\langle r, f\rangle - |N^{-1/2}B^*r|^2]\,dt \tag{3.61}$$

and the optimal response by

$$\bar{y}(t) = \int_{-\infty}^{t} U_L(t, s)[f(s) - K(s)r(s)]\,ds. \tag{3.62}$$

Remark 3.2. The conclusion of Corollary 3.6 is still valid even if we replace (H4) by a weaker condition:

(H4′) There exist numbers ε_1, $\varepsilon_2 \in (0,1)$ such that $\sigma(U_L(t+\theta,t)) \subset \{\lambda \in C: |\lambda| \leqq 1-\varepsilon_1\} \cup \{\lambda \in C: |\lambda| \geqq 1+\varepsilon_2\}$, for all $t \in \mathbb{R}^1$.

Note that the θ-periodic solution Q_∞ still exists and U_L is well defined. If (H4′) holds, we set

$$\Pi_-(t) = \frac{1}{2\pi i} \int_{C_1} R(\lambda, U_L(t+\theta, t))\, d\lambda, \qquad t \in \mathbb{R}^1, \tag{3.63}$$

$$\Pi_+(t) = I - \Pi_-(t), \qquad t \in \mathbb{R}^1 \tag{3.64}$$

where C_1 is the unit circle in the complex plane. Then (3.55) and (3.59) have unique almost periodic solutions given by

$$\begin{aligned} \bar{y}(t) = &\int_{-\infty}^{t} U_L(t,s)\Pi_-(s)[f(s) - K(s)r(s)]\, ds \\ &- \int_t^{\infty} U_L(t,s)\Pi_+(s)[f(s) - K(s)r(s)]\, ds, \end{aligned} \tag{3.65}$$

$$\begin{aligned} r(t) = &\int_t^{\infty} U_L^*(t,s)\Pi_+^*(s)Q_\infty(s)f(s)\, ds \\ &- \int_{-\infty}^{t} U_L^*(t,s)\Pi_-^*(s)Q_\infty(s)f(s)\, ds. \end{aligned} \tag{3.66}$$

This can be proved by using estimates

$$|U_L(t,s)\Pi_-(s)| \leqq M_-\, e^{-\omega_-(t-s)}, \quad t \geqq s \quad \text{for some } M_- > 0 \text{ and } \omega_- > 0, \tag{3.67}$$

$$|U_L(t,s)\Pi_+(s)| \leqq M_+\, e^{\omega_+(t-s)}, \quad t \leqq s \quad \text{for some } M_+ > 0 \text{ and } \omega_+ > 0. \tag{3.68}$$

For a proof of (3.67) and (3.68) see [23] when $A(t)$ has a special form $A(t) = A + \tilde{L}(t)$ with $\tilde{L}(t)$ dominated by A, and see [29] in the general case.

4. Optimal quadratic control in the stochastic case.

4.1. Quadratic control under complete observation. We can "stochasticize" all results in § 3. Let $(\Omega, F, F_t, -\infty < t < \infty, P)$ be a stochastic basis and let (W_i), $i = 1, 2, \cdots, N_0$ and W be independent Wiener processes in $\mathbb{R}^1$ and H (Hilbert), respectively, with $\mathrm{Cov}[W(t)] = tW$, $W \in \mathscr{L}^+(H)$ nuclear. We replace (3.1), (3.2), (3.32), (3.33), (3.45), and (3.46), respectively, by

$$dy = [A(t)y + B(t)u]\, dt + G_i(t)y\, dW_i, \qquad y(t_0) = y_0, \tag{4.1}$$

$$J_0(u) = E \int_{t_0}^{\infty} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt, \tag{4.2}$$

$$dy = [A(t)y + B(t)u + f(t)]\, dt + G_i(t)y\, dW_i, \qquad y(t_0) = y_0, \tag{4.3}$$

$$J_1(u) = \overline{\lim_{T\to\infty}} \frac{1}{T} E \int_{t_0}^{t_0+T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt, \tag{4.4}$$

$$dy = [A(t)y + B(t)u + f(t)]\, dt + G_i(t)y\, dW_i + G(t)\, dW, \tag{4.5}$$

$$J_2(u) = \overline{\lim_{T\to\infty}} \frac{1}{2T} E \int_{-T}^{T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt \tag{4.6}$$

where in (4.1), (4.3), and (4.5) $G_i(t)y\,dW_i$ means the sum over $i=1$ to N_0. The sets of admissible controls are given by

$$\mathcal{U}_{\text{ad}}^0 = \{u \in M^2([t_0, \infty)\times\Omega;\ U)\text{: its response has the property } E|y(t)|^2 \to 0 \text{ as } t\to\infty\}, \tag{4.7}$$

$$\mathcal{U}_{\text{ad}}^1 = \left\{u \in M_{\text{loc}}^2([t_0,\infty)\times\Omega;\ U)\text{: } \overline{\lim_{T\to\infty}}\frac{1}{T}\int_{t_0}^{t_0+T}|u(t)|^2\,dt<\infty,\ E|y(t)|^2 \text{ bounded}\right\}, \tag{4.8}$$

$$\mathcal{U}_{\text{ad}}^2 = \left\{u \in M_{\text{loc}}^2((-\infty,\infty)\times\Omega;\ U)\text{: } \overline{\lim_{T\to\infty}}\frac{1}{2T}\int_{-T}^{T}|u(t)|^2\,dt<\infty, \text{ there exists } y(t) \text{ with } E|y(t)|^2 \text{ bounded}\right\} \tag{4.9}$$

where $M^2([t_1, t_2]\times\Omega;\ U)$ is the subspace of $L_W^2([t_1, t_2]\times\Omega;\ U)$, which consists of F_t-adapted processes and M_{loc}^2 means M^2 for any finite intervals. We remark that it is possible to take an infinite-dimensional $\tilde{W}$ in place of (W_i) as in [25].

The Riccati equations (3.4) and (3.28) are replaced by

$$Q'(t)+A^*(t)Q(t)+Q(t)A(t)+G_i(t)Q(t)G_i^*(t) + M^*(t)M(t)-Q(t)B(t)N^{-1}(t)B^*(t)Q(t)=0, \tag{4.10}$$

$$A^*Q+QA+G_iQG_i^*+M^*M-QBN^{-1}B^*Q=0. \tag{4.11}$$

We keep the hypothesis (H1) as it is and replace (H2), (H3), respectively, by

(S2) (H2) together with
(i) $G_i \in C_S(\mathbb{R}^1;\ \mathcal{L}(Y))$, $G \in C_S(\mathbb{R}^1;\ \mathcal{L}(H, Y))$; and
(ii) $\sup_{t\in\mathbb{R}^1}[|G_i(t)|+|G(t)|]<\infty$.

(S3) For each $t_0\in\mathbb{R}^1$ and $y_0\in L^2(\Omega, F_{t_0}, P)$, there exist $u\in M^2([t_0,\infty)\times\Omega;\ U)$ and C_0 such that

$$E\int_{t_0}^{\infty}[|M(t)y|^2+\langle N(t)u, u\rangle\,dt \leqq C_0E|y_0|^2$$

where y is the mild solution of (4.1).

To replace (H4) by a new one we need to recall some definitions. Consider the homogeneous system

$$dy = A(t)y\,dt + G_i(t)y\,dW_i, \qquad y(t_0)=y_0. \tag{4.12}$$

Since (4.12) is linear, we can easily establish a unique mild solution in $C([t_0, T];\ L_2(\Omega;\ Y))$ that is adapted to F_t: namely, the solution of

$$y(t) = U(t, t_0)y_0 + \int_{t_0}^{t} U(t, s)G_i(s)y(s)\,dW_i(s). \tag{4.13}$$

We say that $(A;\ G_i)$ is (*exponentially*) *stable* if the mild solution of (4.12) satisfies

$$E|y(t)|^2 \leqq M_1\,e^{-\omega(t-t_0)}E|y_0|^2 \quad \forall y_0\in L^2(\Omega, F_{t_0}, P), \quad t\geqq t_0$$

for some $M_1\geqq 1$ and $\omega>0$. Let $V(t, s)$: $L^2(\Omega, F_s, P)\to L^2(\Omega, F_t, P)$ be the stochastic fundamental solution [2], [14], [27] so that $y(t)=V(t, t_0)y_0$. Then (A, G_i) is stable if and only if

$$E|V(t, t_0)y_0|^2 \leqq M_1\,e^{-\omega(t-t_0)}E|y_0|^2, \qquad y_0\in L^2(\Omega, F_{t_0}, P).$$

We say that $(A, B; G_i)$ is *stabilizable* if there exists a $K \in C_s(\mathbb{R}^1; \mathscr{L}(Y, U))$ bounded such that $(A-BK; G_i)$ is stable. Hypothesis (S3) is fulfilled if $(A, B; G_i)$ is stabilizable. Let $D \in C_s(\mathbb{R}^1; \mathscr{L}(Y))$ be bounded. We say that $(A, D; G_i)$ is *detectable* if there exists a $K_1 \in C_S(\mathbb{R}^1; \mathscr{L}(Y))$ bounded such that $(A-K_1D; G_i)$ is stable.

Now we replace (H4) and (H5), respectively, by

(S4) $(A, M; G_i)$ is detectable,

and

(S5) (H5) and $G_i(t+\theta)=G_i(t)$, $G(t+\theta)=G(t)$, $t \in \mathbb{R}^1$.

Now we can stochasticize almost all results in § 3 under our new hypotheses, but below we will give only main results.

THEOREM 4.1. (i) *Assume* (H1) *and* (S2). *Then a nonnegative bounded solution to* (4.10) *exists if and only if* (S3) *holds.*

(ii) *Assume* (H1), (S2), *and* (S4). *Then any bounded nonnegative solution of* (4.10) *is stable, i.e.,* $(A-BN^{-1}B^*Q; G_i)$ *is stable. Hence the Riccati equation* (4.10) *has at most one bounded nonnegative solution in* $\mathscr{L}^+(Y)$.

PROPOSITION 4.1. (i) *Assume* (H1), (S2), (S3), *and* (S5). *Then the minimal solution* Q_∞ *of* (4.10) *is* θ*-periodic. If, further,* (S4) *holds, then* Q_∞ *is the unique* θ*-periodic solution to* (4.10).

(ii) *Suppose all operators are constant. Then* Q_∞ *is constant and is the minimal solution of the Riccati equation* (4.11). *If* $(A, M; G_i)$ *is detectable* [15], *then* Q_∞ *is the unique solution of* (4.11) *in* $\mathscr{L}^+(Y)$ *and* $(A-BN^{-1}BQ_\infty; G_i)$ *is stable.*

Remark 4.1. Results similar to Propositions 3.1, 3.2, 3.5 and Corollaries 3.1–3.3 are also valid.

Now we need a result similar to Lemmas 3.1 and 3.2.

LEMMA 4.1. *Assume* (H1), (S2), *and* (S3). *Let* Q_∞ *be the minimal nonnegative solution of* (4.10). *Let* $u \in M^2([t_1, t_2]\times\Omega; U)$ *and let* y, r *be any mild solutions of* (4.5) *and*

$$r' + L^*r + Q_\infty f = 0, \qquad L^* = A - BN^{-1}B^*Q_\infty, \tag{4.14}$$

respectively, on $[t_1, t_2]$. *Then*

$$\begin{aligned} E\int_{t_1}^{t_2} [|M(t)y|^2 + \langle N(t)u, u\rangle]\,dt + E\langle Q_\infty(t_2)y(t_2), y(t_2)\rangle + 2E\langle r(t_2), y(t_2)\rangle \\ = E\int_{t_1}^{t_2} |N^{1/2}[u + N^{-1}B^*(Q_\infty y + r)]|^2\,dt \\ + \int_{t_1}^{t_2} [2\langle r, f\rangle - |N^{-1/2}B^*r|^2 + \operatorname{tr} GWG^*Q_\infty]\,dt \\ + E\langle Q_\infty(t_1)y(t_1), y(t_1)\rangle + 2E\langle r(t_1), y(t_1)\rangle. \end{aligned} \tag{4.15}$$

Proof. We apply Itô's formula to $\langle Q_n(t)y_n(t), y_n(t)\rangle + 2\langle r_n(t), y_n(t)\rangle$, where Q_n, y_n, r_n are approximations to Q_∞, y, r given by (4.10), (4.5), and (4.14). Then we rearrange terms, take expectations, and finally pass to the limit $n \to \infty$. □

Now we can solve our three problems immediately.

THEOREM 4.2. *Assume* (H1), (S2)–(S4) *and consider the control problems* (4.1), (4.2), (4.7). *The optimal control is given by the feedback law*

$$\bar{u} = -N^{-1}B^*Q_\infty y \tag{4.16}$$

where Q_∞ is the unique bounded nonnegative solution of (4.10) *and the optimal cost is*

$$J_0(\bar{u}) = E\langle Q_\infty(t_0)y_0, y_0\rangle. \tag{4.17}$$

The optimal closed-loop system is stable.

If (S5) *holds, then Q_∞ is θ-periodic. If all operators in* (4.1), (4.2) *are constant, then Q_∞ is constant.*

Proof. We set $f = r = 0$, $t_1 = t_0$, $t_2 = \infty$, and $G = 0$ in (4.15). □

THEOREM 4.3. *Assume* (H1), (S2)-(S4) *and consider the control problems* (4.3), (4.4), (4.8). *The optimal control is given by the feedback law*

$$\bar{u} = -N^{-1}B^*(Q_\infty \bar{y} + r) \tag{4.18}$$

where Q_∞ is the unique bounded nonnegative solution of (4.10) *and r is the unique bounded solution of*

$$r' + L^*r + Q_\infty f = 0, \qquad L = A - BN^{-1}BQ_\infty, \tag{4.19}$$

given by

$$r(t) = \int_t^\infty U_L^*(s, t)Q_\infty(s)f(s)\, ds, \qquad t \geqq t_0. \tag{4.20}$$

The optimal cost is

$$J_1(u) = \overline{\lim_{T\to\infty}} \frac{1}{T}\int_{t_0}^{t_0+T} [2\langle r, f\rangle - |N^{-1/2}B^*r|^2 + \operatorname{tr} GWG^*Q_\infty]\, dt. \tag{4.21}$$

If, further, (S5) *holds, and f is θ-periodic, then Q_∞, r are θ-periodic and*

$$J_1(\bar{u}) = \frac{1}{\theta}\int_{t_0}^{t_0+\theta} [2\langle r, f\rangle - |N^{1/2}B^*r|^2 + \operatorname{tr} GWG^*Q]\, dt. \tag{4.22}$$

THEOREM 4.4. *Assume* (H1), (S2)-(S4) *and consider the control problem* (4.5), (4.6), (4.9). *Then the optimal control is given by*

$$\bar{u} = -N^{-1}B^*(Q_\infty \bar{y} + r) \tag{4.23}$$

where Q_∞ is the unique bounded nonnegative solution of (4.10) *on $\mathbb{R}^1$ and r is the unique bounded solution on $\mathbb{R}^1$ of*

$$r' + L^*r + Q_\infty f = 0, \qquad L = A - BN^{-1}B^*Q \tag{4.24}$$

given by

$$r(t) = \int_t^\infty U_L^*(s, t)Q_\infty(s)f(s)\, ds, \qquad t \in \mathbb{R}^1. \tag{4.25}$$

The optimal cost is given by

$$J_2(\bar{u}) = \overline{\lim_{T\to\infty}} \frac{1}{2T}\int_{-T}^{T} [2\langle r, f\rangle - |N^{-1/2}B^*r|^2 + \operatorname{tr} GWG^*Q_\infty]\, dt \tag{4.26}$$

and the optimal closed-loop system by

$$d\bar{y} = [(A - BN^{-1}B^*Q_\infty)\bar{y} + f - BN^{-1}B^*r]\, dt + G_i\bar{y}\, dw_i + G\, dw. \tag{4.27}$$

It has a unique bounded solution

$$\begin{aligned}\bar{y}(t) = &\int_{-\infty}^t V_Q(t, s)[f(s) - B(s)N^{-1}(s)B^*(s)r(s)]\, ds \\ &+ \int_{-\infty}^t V_Q(t, s)G(s)\, dw(s)\end{aligned} \tag{4.28}$$

where $V_Q(t, s)$ is the stochastic fundamental solution associated with the homogeneous part of (4.27). *$\bar{y}(t)$ is exponentially asymptotically stable, i.e., any solution $y(t)$ of* (4.27) *with $y(0) = y_0$ satisfies: $y(t) - \bar{y}(t) \to 0$ exponentially in mean square as $t \to \infty$.*

If, further, (S5) *holds, and f is θ-periodic, then Q_∞, r are θ-periodic and*

$$J_2(\bar{u}) = \frac{1}{\theta} \int_0^\theta [2\langle r, f\rangle - |N^{-1/2}B^*r|^2 + \operatorname{tr} GWG^*Q_\infty]\, dt \tag{4.29}$$

and $\bar{y}(t)$, given by (4.28), *is the unique θ-periodic solution of* (4.27).

Finally, we consider almost periodic controls. We say that a stochastic process $z(t)$ is (*weakly*) *almost periodic* in Z if $Ez(t)$ and $\operatorname{cov}[z(t)]k$, $k \in Z$ are almost periodic. We assume that all operators except G are θ-periodic and that $G(t)h$, $h \in H$, f are almost periodic. We wish to minimize

$$J_{\mathrm{ap}}(u) = \lim_{T\to\infty} \frac{1}{2T} E \int_{-T}^{T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt \tag{4.30}$$

subject to (4.5) over

$$\begin{aligned} \mathcal{U}_{\mathrm{ap}} = \{u:\ &\text{adapted to } F_t, \text{ almost periodic such that there exists} \\ &\text{a mild solution } y \text{ of (4.5) almost periodic}\}. \end{aligned} \tag{4.31}$$

THEOREM 4.5. *Assume* (H1), (S2)-(S4), *and* (S5) *except G. Assume that $G(t)h$, for all $h \in H$ and f are almost periodic. Consider the control problem* (4.5), (4.30), (4.31). *Then the optimal control is given by the feedback law*

$$\bar{u} = -N^{-1}B^*(Q_\infty \bar{y} + r) \tag{4.32}$$

where Q_∞ is the unique stable θ-periodic solution of (4.10) *and r is the unique almost periodic solution of*

$$r' + L^*r + Q_\infty f = 0 \tag{4.33}$$

given by

$$r(t) = \int_t^\infty U_L^*(s, t) Q_\infty(s) f(s)\, ds. \tag{4.34}$$

The optimal cost is

$$J_{\mathrm{ap}}(\bar{u}) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} [2\langle r, f\rangle - |N^{-1/2}B^*r|^2 + \operatorname{tr} GWG^*Q_\infty]\, dt. \tag{4.35}$$

The optimal closed-loop system is given by (4.37) *and its unique solution is given by* (4.28).

4.2. Quadratic control under partial observation. Consider a special case of (4.3) and its observation

$$dy = [A(t)y + B(t)u + f(t)]\, dt + G(t)\, dw, \qquad y(t_0) = y_0, \tag{4.36}$$

$$dz = C(t)y\, dt + V(t)\, dv, \qquad z(t_0) = 0 \tag{4.37}$$

where $C(t) \in C_b(\mathbb{R}^1; L(Y, \mathbb{R}^m))$, $V(t) \in C_b(\mathbb{R}^1; \mathbb{R}^{m\times n})$, nonsingular, v is an m-dimensional Wiener process, $y_0 \in L^2(\Omega, F_{t_0}, P)$ is Gaussian with mean $\bar{y}_0$ and covariance P_0, and y_0, w, v are independent. We assume (H1), (S2)-(S4) and wish to minimize

$$J_1(u) = \overline{\lim_{T\to\infty}} \frac{1}{T} E \int_{t_0}^{t_0+T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\, dt \tag{4.38}$$

over all controls u that are adapted to $\sigma\{z(s), t_0 \leqq s \leqq t\}$. We will define the set of admissible controls later. We now recall filtering results of the system:

$$dy = A(t)y\,dt + G(t)\,dw, \qquad y(t_0) = y_0, \tag{4.39}$$

$$dz = C(t)y\,dt + V(t)\,dw, \qquad z(t_0) = 0. \tag{4.40}$$

The optimal filter $\hat{y}(t)$ of $y(t)$ given $Z_t = \sigma\{z(s), t_0 \leqq s \leqq t\}$ is the projection of $y(t)$ onto $L^2(\Omega, Z_t, P)$ [20], [26] and is given by the mild solution of

$$d\hat{y} = A(t)\hat{y}\,dt + P(t)C^*(t)[VV^*(t)]^{-1}\,d\eta, \qquad \hat{y}(t_0) = \bar{y}_0 \tag{4.41}$$

where η is the innovation process given by

$$d\eta = dz - C(t)\hat{y}\,dt \tag{4.42}$$

and $P(t)$, the covariance of the error process $e = y - \hat{y}$, is the mild solution of

$$\begin{aligned} &\text{(a)} \quad P'(t) - A(t)P(t) - P(t)A^*(t) - G(t)WG^*(t) \\ &\qquad + P(t)C^*(t)[VV^*(t)]^{-1}C(t)P(t) = 0, \\ &\text{(b)} \quad P(t_0) = P_0. \end{aligned} \tag{4.43}$$

Following [3], [6], [9], and [17] we define the set of admissible controls

$$\begin{aligned} \mathcal{U}_{\text{pad}} = \Bigg\{ & u \in M^2_{\text{loc}}([t_0, \infty) \times \Omega;\, U)\colon \lim_{T\to\infty} \frac{1}{T} E \int_{t_0}^{t+T_0} |u(t)|^2\,dt \\ & < \infty,\ u(t) \in L^2(\Omega, H_t, P;\, U) \cap L^2(\Omega, Z_t, P;\, U) \\ & \qquad \text{a.e. } t \text{ and } E|y(t)|^2 \text{ is bounded} \Bigg\}, \end{aligned} \tag{4.44}$$

where $H_t = \sigma\{\eta(s), t_0 \leqq s \leqq t\}$.

Now let u be an admissible control and define $\hat{y}$ by

$$d\hat{y} = [A(t)\hat{y} + B(t)u + f(t)]\,dt + P(t)C^*(t)[VV^*(t)]^{-1}\,d\eta, \qquad \hat{y}(0) = \bar{y}_0. \tag{4.45}$$

Then it is well known [3], [6], [9], [17] that

$$\begin{aligned} & E \int_{t_0}^{t_0+T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\,dt \\ & \qquad = E \int_{t_0}^{t_0+T} [|M(t)y|^2 + \langle N(t)u, u\rangle]\,dt + \int_{t_0}^{t_0+T} \operatorname{tr} M(t)P(t)M^*(t)\,dt \end{aligned} \tag{4.46}$$

where y is the response of (4.36). To make our problem nontrivial we assume:

(S6) (a) (A^*, C) is stabilizable,
(b) $(A^*, W^{1/2}G^*)$ is detectable.

PROPOSITION 4.2. *Assume* (H1) *and* (S6). *Then there exists a unique bounded stable solution* P_∞ *to* (4.43a). *The solution* $P(t)$ *of* (4.43) *is bounded on* $[t_0, \infty)$ *for any* $P_0 \geqq 0$. *If, further,* $A(t)$, $C(t)$, *and* $V(t)$ *are* θ*-periodic, then* $P_\infty(t)$ *is* θ*-periodic and* $P(t + n\theta) \to P_\infty(t)$ *strongly for any* $t \geqq T_0$ *as* $n \to \infty$.

Proof. The Riccati equation (4.4) is dual to (3.4). Hence the assertions follow from Propositions 3.2–3.5. □

Remark 4.2. We may replace (S6)(a) by a condition similar to (H3) for the dual control problem. Note that under (S6), P is bounded and

$$\overline{\lim_{T\to\infty}}\frac{1}{T}\int_{t_0}^{t_0+T}\operatorname{tr} M(t)P(t)M^*(t)\,dt<\infty.$$

Now consider an auxiliary problem of minimizing

$$\hat{J}_1(u)=\overline{\lim_{T\to\infty}}\frac{1}{T}\int_{t_0}^{t_0+T}[|M(t)y|^2+\langle N(t)u,u\rangle]\,dt$$

subject to (4.41) over

$$\hat{\mathcal{U}}_{\text{ad}}=\Big\{u:\overline{\lim_{T\to\infty}}\frac{1}{T}E\int_{t_0}^{t_0+T}|u(t)|^2\,dt$$

$$<\infty,\ u \text{ adapted to } H_t \text{ such that } E|\hat{y}(t)|^2 \text{ is bounded}\Big\}.$$

Assume (H1), (H3), (H4), (S2), and (S6). Then in view of Theorem 4.1 the optimal control is given by

$$\bar{u}=-N^{-1}B^*(Q_\infty\hat{y}+r)$$

where Q_∞, r are given as in Theorem 3.4 and

$$J_1(\bar{u})=\overline{\lim_{T\to\infty}}\frac{1}{T}\int_{t_0}^{t_0+T}[2\langle r,f\rangle-|N^{-1/2}B^*r|^2+\operatorname{tr} PC^*[VV^*]^{-1}CPQ_\infty]\,dt.$$

It is well known that this $\bar{u}$ lies in $\mathcal{U}_{\text{pad}}$ [6], [9]. Then we have Theorem 4.6.

THEOREM 4.6. *Assume* (H1), (H3), (H4), (S2), *and* (S6) *and consider the control problem* (4.36)-(4.38), (4.44). *Then the optimal control is given by*

$$\bar{u}=-N^{-1}B^*(Q_\infty\hat{y}+r)$$

where Q_∞ is the unique bounded stable solution of (3.4) *and r is the unique bounded solution of* (3.49) *given by* (3.50). *The optimal cost is given by*

$$J_1(\bar{u})=\overline{\lim_{T\to\infty}}\frac{1}{T}\int_{t_0}^{t_0+T}[2\langle r,f\rangle-|N^{-1/2}B^*r|^2$$

$$+\operatorname{tr} MPM^*+\operatorname{tr} PC^*[VV^*]^{-1}CPQ_\infty]\,dt.$$

If, further, $f(t)$, $C(t)$, and $V(t)$ are θ-periodic and (S5) *holds, then Q_∞, r are θ-periodic and*

$$J_1(\bar{u})=\frac{1}{\theta}\int_0^\theta[2\langle r,f\rangle-|N^{-1/2}B^*r|^2+\operatorname{tr} MP_\infty M^*+\operatorname{tr} P_\infty C^*[VV^*]^{-1}CP_\infty Q_\infty]\,dt$$

where P_∞ is the unique θ-periodic solution of (4.43a).

We may also consider two-sided average cost as J_2 in § 4.1 although the problem becomes a little artificial. We replace the initial conditions of (4.36), (4.37) by $y(-T)=y_0$, $z(-T)=0$ and minimize

$$J_2(u)=\overline{\lim_{T\to\infty}}\frac{1}{2T}\int_{-T}^{T}[|M(t)y|^2+\langle N(t)u,u\rangle]\,dt.$$

We allow only for feedback controls on the filtered process $\hat{y}$ of the form

$$u=-K(t)\hat{y}+h(t)$$

where $K(t)\in C_S(\mathbb{R}^1;\mathcal{L}(Y,U))$ is bounded and $h\in C_b(\mathbb{R}^1;U)$.

THEOREM 4.7. *Assume* (H1), (H3), (H4), (S2), *and* (S6). *The optimal control is given by*

$$\bar{u} = -N^{-1}B^*(Q_\infty \bar{y} + r)$$

and

$$J_2(\bar{u}) = \overline{\lim_{T\to\infty}} \frac{1}{T}\int_{-T} [2\langle r, f\rangle - |N^{-1/2}B^*r|^2 + \operatorname{tr} MP_\infty M^* + \operatorname{tr} P_\infty C^*[VV^*]^{-1}CP_\infty Q_\infty]\, dt$$

where Q_∞, P_∞, *and* r *are unique bounded solutions of* (3.4), (4.43a), *and* (3.49), *respectively. If* f, C, V *are* θ*-periodic and* (S5) *holds, then* Q_∞, P_∞, r *are* θ*-periodic and*

$$J_2(\bar{u}) = \frac{1}{\theta}\int_0^\theta [2\langle r, f\rangle - |N^{-1/2}B^*r|^2 + \operatorname{tr} MP_\infty M^* + \operatorname{tr} P_\infty C^*[VV^*]^{-1}CP_\infty Q_\infty]\, dt.$$

Remark 4.1. We are not able to prove the existence of almost periodic P if the coefficients of (4.43) are almost periodic.

5. An example. Consider the system:

$$\begin{aligned}
&\frac{\partial y}{\partial t} = \sum_{i,j=1}^n \frac{\partial}{\partial x_j} a_{ij}(t,x)\frac{\partial y}{\partial x_i} + \sum_{i=1}^n b_i(t,x)\frac{\partial y}{\partial x_i} \\
&\qquad + c(t,x)y + f(t,x) + u(t,x), \qquad (t,x)\in \mathbb{R}\times\Omega, \quad a_{ij} = a_{ji}, \\
&y(t,x) = 0, \qquad (t,x)\in \mathbb{R}\times\partial\Omega, \\
&y(0,x) = y_0(x), \qquad x\in\Omega
\end{aligned} \tag{5.1}$$

where a_{ij}, b_i, c, f, n are real functions from $\mathbb{R}\times\bar{\Omega}$ to $\mathbb{R}$ and Ω is a bounded set in $\mathbb{R}^n$ with smooth boundary $\partial\Omega$.

We assume the following:

(5.2)
(i) a_{ij}, b, c, and f are continuous and bounded with their first derivatives with respect to $x\in\bar{\Omega}$ and $t\in\mathbb{R}$.
(ii) There exists $\nu > 0$ such that

$$\sum_{i,j=1}^n a_{ij}(t,x)\xi_i\zeta_j \geqq \nu|\zeta|^2, \qquad \xi = (\zeta_1, \cdots, \zeta_n)\in\mathbb{R}^n, \quad x\in\bar{\Omega}, \quad t\in\mathbb{R}.$$

We set $Y = L^2(\Omega)$ and denote by $A(t)$ the linear operator in Y

$$\begin{aligned}
&A(t)y = \sum_{i,j=1}^n \frac{\partial}{\partial x_j} a_{ij}(t,x)\frac{\partial y}{\partial x_i} + \sum_{i=1}^n b_i(t,x)\frac{\partial y}{\partial x_i} + c(t,x)y, \\
&D(A(t)) = H^2(\Omega)\cap H_0^1(\Omega).
\end{aligned} \tag{5.3}$$

Then hypothesis (H1)(i) is fulfilled (see [36]) and (H1)(ii) is easily checked (see, for instance, Tanabe [36]). Moreover, the adjoint operator $A^*(t)$ is given by

$$\begin{aligned}
&A^*(t)y = \sum_{i,j=1}^n \frac{\partial}{\partial x_j} a_{ij}(t,x)\frac{\partial y}{\partial x_i} - \sum_{i=1}^n \frac{\partial}{\partial x_i}(b_i(t,x)y) + c(t,x)y, \\
&D(A^*(t)) = H^2(\Omega)\cap H_0^1(\Omega)
\end{aligned}$$

so that (H1)(ii) holds.

Consider the quadratic control problem. Minimize

$$J_0(u) = \int_0^\infty dt \int_\Omega (|y(t,x)|^2 + |u(t,x)|^2)\, dx \tag{5.4}$$

subject to (5.1) over the set of admissible controls

$$U_{ad}^0 = \{u \in L^2(0, \infty; Y); \text{ the corresponding mild solution to (5.1), } y(t) \to 0 \text{ as } t \to \infty\}. \tag{5.5}$$

We take $U = Y = L^2(\Omega)$, $B = N = M = I$. Then (H2)-(H4) are fulfilled. Thus the Riccati equation (3.4) has a unique bounded solution and the hypotheses of Theorem 3.2 are fulfilled. Then there exists an optimal feedback control for the infinite horizon problem (5.4), (5.5).

Consider now the stochastic system:

$$\begin{aligned} dy = &\Bigg(\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j} a_{ij}(t,x) \frac{\partial y}{\partial x_i} \\ &+ \sum_{i=1}^{n} b_i(t,x) \frac{\partial y}{\partial x_i} + c(t,x)y + u(t,x) + f(t,x) \Bigg) dt \\ &+ \sum_{i=1}^{n} g_i(t,x) y \, dw_i + g(t,x) \, dw, \qquad (t,x) \in [0, +\infty) \times \bar{\Omega}, \\ &y(t,x) = 0, \qquad (t,x) \in [0, +\infty) \times \partial\Omega, \\ &y(0,x) = y_0(x), \qquad x \in \bar{\Omega} \end{aligned} \tag{5.6}$$

where g, g_i are also continuous and bounded with their first derivatives.

Consider the problem. Minimize

$$J_0(u) = E \int_0^\infty dt \int_\Omega (|y(t,x)|^2 + |u(t,x)|^2) \, dt \tag{5.7}$$

over all $u \in U_{ad}^0$ defined by (4.7) where y is the solution of (5.6). Now we can apply Theorem 4.2 and so there exists a feedback for problem (3.9).

REFERENCES

[1] L. AMERIO AND G. PROUSE, *Almost Periodic Functions and Functional Equations*, Van Nostrand, New York, 1971.

[2] L. ARNOLD, *Stochastic Differential Equations: Theory and Applications*, John Wiley, New York, 1974.

[3] A. V. BALAKRISHNAN, *Stochastic Differential Systems* I, Lecture Notes in Economics Math. Systems, Springer-Verlag, Berlin, 1973.

[4] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman London, 1983.

[5] A. BENSOUSSAN, *Filtrage Optimal des Systemes Linéaires*, Dunod, Paris, 1971.

[6] A. BENSOUSSAN AND M. VIOT, *Optimal control of stochastic linear distributed parameter systems*, SIAM J. Control, 13 (1975), pp. 904-926.

[7] S. BITTANTI, A. LOCATELLI, AND C. MAFFEZZONI, *Periodic optimization under small perturbations*, in Periodic Optimization Vol. II, A. Marzollo, ed., Springer-Verlag, New York, 1972, pp. 183-231.

[8] F. COLONIUS, *Optimal periodic control*, Report 140, Forschungsschwerpunkt Dynamische Systeme, Universität Bremen, Bremen, West Germany, October 1985.

[9] F. CURTAIN AND A. ICHIKAWA, *The separation principle for stochastic evolution equations*, SIAM J. Control Optim., 15 (1977), pp. 367-383.

[10] F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control Information Science 8, Springer-Verlag, New York, 1978.

[11] G. DA PRATO, *Quelques résultats d'existence, unicité et regularité pour un problème de la théorie du contrôle*, J. Math. Pure Appl., 52 (1973), pp. 353-375.

[12] ———, *Periodic Solutions of an infinite dimensional Riccati equation*, in Proc. 12th IFIP Conference on Systems Modelling and Optimizations, Budapest, Hungary, 1985, Lecture Notes in Control and Information Sciences 84, Springer-Verlag, Berlin, 1985, pp. 714-722.

[13] G. DA PRATO, *Equations aux derivées partielles stochastiques et applications*, R.I. No. 148, Centre de Mathématiques Appliquées, Ecole Polytechnique, France, 1986.
[14] ———, *Synthesis of optimal control for an infinite dimensional periodic problem*, SIAM J. Control Optim., 25 (1987), pp. 706-714.
[15] G. DA PRATO AND A. ICHIKAWA, *Stability and quadratic control for linear stochastic equations with unbounded coefficients*, Boll. Un. Mat. Ital. B(6), 4 (1985), pp. 987-1001.
[16] ———, *Optimal control of linear systems with almost periodic inputs*, SIAM J. Control Optim., 25 (1987), pp. 1007-1019.
[17] ———, *Quadratic control for linear periodic systems*, Appl. Math. Optim., 18 (1988), pp. 39-66.
[18] P. DATKO, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1974), pp. 428-455.
[19] ———, *Some nonautonomous control problems with quadratic cost*, J. Differential Equations, 21 (1976), pp. 231-262.
[20] M. H. A. DAVIS, *Linear Estimation and Stochastic Control*, Chapman and Halls, London, 1977.
[21] A. M. FINK, *Almost Periodic Differential Equations*, Lecture Notes in Mathematics 377, Springer-Verlag, Berlin, 1974.
[22] C. J. HARRIS AND J. F. MILES, *Stability of Linear Systems: Some Aspects of Kinematic Similarity*, Academic Press, London, 1980.
[23] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics 840, Springer-Verlag, Berlin, 1981.
[24] A. ICHIKAWA, *Optimal control of a linear stochastic evolution equation with state and control dependent noise*, in Proc. IMA Conference, Recent Theoretical Developments in Control, Leicester, UK, Academic Press, London, 1978, pp. 383-401.
[25] ———, *Dynamic programming approach to stochastic evolution equations*, SIAM J. Control Optim., 17 (1979), pp. 153-174.
[26] ———, *Filtering and control of stochastic differential equations with unbounded coefficients*, Stochastic Anal. Appl., 4 (1986), pp. 187-212.
[27] ———, *Bounded solutions and periodic solutions of a linear stochastic evolution equation, 5th Japan-USSR Symposium on Probability Theory*, Kyoto, July 1986, Lecture Notes in Mathematics 1299, Springer-Verlag, Berlin, 1988, pp. 124-130.
[28] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.
[29] A. LUNARDI, *Bounded solutions of linear periodic abstract parabolic equations*, Proc. Royal Soc. Edinburgh Sect. A, 110 (1988), pp. 135-159.
[30] T. MOROZAN, *Periodic solutions of affine stochastic differential equations*, Stochastic Anal. Appl., 4 (1986), pp. 87-110.
[31] T. NISHIMURA AND H. KANO, *Periodic solutions of matrix Riccati equations with detectability and stabilizability*, Internat. J. Control, 29 (1979), pp. 471-487.
[32] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[33] A. J. PRITCHARD AND J. ZABCZYK, *Stability and stabilizability of infinite-dimensional systems*, SIAM Rev., 23 (1981), pp. 25-52.
[34] M. A. SHAYMAN, *Phase portrait of the matrix Riccati equation*, SIAM J. Control Optim., 24 (1982), pp. 1-64.
[35] J. L. SPEYER AND R. T. EVANS, *A second variational theory for optimal periodic processes*, IEEE Trans. Automat. Control, 29 (1984), pp. 138-147.
[36] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.
[37] Y. V. VENKATESH, *Energy Methods in Time-Varying System Stability and Instability Analyses*, Lecture Notes in Physics 68, Springer-Verlag, Berlin, 1977.
[38] W. M. WONHAM, *Optimal stationary control of a linear system with state-dependent noise*, SIAM J. Control, 5 (1967), pp. 486-500.
[39] ———, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681-697.
[40] ———, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, Vol. 2, A. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131-212.
[41] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optim., 2 (1976), pp. 251-258.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 382-394, March 1990

007

APPROXIMATIONS TO STOCHASTIC PROGRAMS WITH COMPLETE RECOURSE*

RIHO LEPP†

Abstract. A two-stage stochastic convex program with relatively complete recourse in L^∞-space is replaced by a mathematical programming problem in a finite-dimensional space. Using discrete convergence of mappings, conditions are presented that guarantee the convergence of the sequence of solutions of approximate problems to the solution of the initial problem.

Key words. stochastic programming, discrete approximation, relatively complete recourse

AMS(MOS) subject classifications. 90C15, 90C31, 65J15

1. Introduction. Let C_1 and C_2 be nonempty, closed convex sets in R^r and R^v, respectively, and let (S, Σ, σ) be a probability space with $S \subset R^k$ and Σ the Borel sigma-field on S. Let f_{1j} be a finite convex function on R^r for $j = 0, 1, \cdots, l_1$ and let $f_{2j}(s, \cdot, \cdot)$ be a finite convex function on $R^r \times R^v$ for $j = 0, 1, \cdots, l_2$ and for almost all $s \in S$.

Consider the stochastic programming problem with recourse: minimize the function

$$f_{10}(x) + \int_S Q(s, x)\sigma(ds) \tag{1.1}$$

over all $x \in R^r$ satisfying

$$x \in C_1 \quad \text{and} \quad f_{1j}(x) \leqq 0, \qquad j = 1, \cdots, l_1, \tag{1.2}$$

where the function $Q(s, x)$ is defined as the value of an "inner" (second stage) subproblem:

$$Q(s, x) = \inf_{y \in C_2} \{f_{20}(s, x, y) | f_{2j}(s, x, y) \leqq 0, \qquad j = 1, \cdots, l_2\}. \tag{1.3}$$

Properties of recourse function $\mathcal{Q}(x)$,

$$\mathcal{Q}(x) = \int_S Q(s, x)\sigma(ds) \tag{1.4}$$

(continuity, convexity) are widely known (see, e.g., [10]). In spite of these promising properties, stochastic programming problems with recourse generally cannot be solved by known methods of mathematical programming because the numerical evaluation of $\mathcal{Q}(x)$ and its gradient is an extremely complicated problem.

Several authors have examined the effect of perturbations in problem data on the optimal value and the solution of the problem (1.1)-(1.3) (see [4], [8], [23] and the cited literature in these references). These results are based, generally speaking, on properties of the value function $Q(s, x)$ and on the weak convergence of a sequence

* Received by the editors August 31, 1987; accepted for publication (in revised form) March 13, 1989.
† Institute of Cybernetics, Estonian Academy of Sciences, Akadeemia 21, SU-200108 Tallinn, Union of Soviet Socialist Republics.

of discrete probability measures $\{\sigma_n\}$ to measure σ (see, e.g., [3]). Sequence $\{\sigma_n\}$ is determined by partitions $\{\mathcal{S}_n\}$ of S, $\mathcal{S}_n = \{S_{1n}, \cdots, S_{nn}\}$, with properties

(1) $\bigcup_{i=1}^{n} S_{in} = S$;

(2) $S_{in} \cap S_{jn} = \emptyset$, $i \neq j$;

(3) $\mathcal{S}_n \subset \mathcal{S}_{n+1}$;

(4) $\max_{1 \leqq i \leqq n} \sigma(S_{in}) \to 0$, $n \to \infty$.

The partition $\{\mathcal{S}_n\}$ is used in order to get lower and upper bounds for the solution of (1.1)-(1.3). Construction of these bounds rely on the Jensen's inequality and Edmundson-Madansky bound, respectively. In such a partition it is necessary to find the discretization points in the form of conditional mean [4]:

$$\bar{s}_{in} = E\{s \mid S_{in}\}. \tag{1.5}$$

Then the problem reads as follows: minimize the function

$$f_{10}(x) + \sum_{i=1}^{n} Q(\bar{s}_{in}, x) p_{in} \tag{1.6}$$

over all $x \in R^r$ satisfying

$$x \in C_1 \quad \text{and} \quad f_{1j}(x) \leqq 0, \qquad j = 1, \cdots, l_1, \tag{1.7}$$

where

$$Q(\bar{s}_{in}, x) = \inf_{y \in C_2} \{f_{20}(\bar{s}_{in}, x, y) \mid f_{2j}(\bar{s}_{in}, x, y) \leqq 0, \qquad j = 1, \cdots, l_2\} \tag{1.8}$$

and $p_{in} = \sigma(S_{in})$.

Equivalently, if we are able to replace infimum by the minimum for all $\bar{s}_{in}$, then instead of (1.6)-(1.8) we get the following extremum problem (see, e.g., [10]) with activities $y_i \in R^v$, $i = 1, \cdots, n$,: minimize over x and y_i, $i = 1, \cdots, n$, the function

$$f_{10}(x) + \sum_{i=1}^{n} f_{20}(\bar{s}_{in}, x, y_i) p_{in} \tag{1.9}$$

satisfying conditions

$$x \in C_1, \qquad f_{1j}(x) \leqq 0, \qquad j = 1, \cdots, l_1; \tag{1.10}$$

$$y_i \in C_2, \qquad f_{2j}(s_{in}, x, y_i) \leqq 0, \qquad j = 1, \cdots, l_2, \quad i = 1, \cdots, n. \tag{1.11}$$

Roughly speaking, in [4], [9], [23] the approximation scheme is realized via the approximation of an integral by Lebesgue sums.

The aim of this paper is to investigate the behaviour of the sequence of second stage solutions $\{\bar{y}_{in}\}$ of the problems (1.8) as $n \to \infty$. We show that it converges in a certain (weak* discrete) sense to an essentially bounded function $\bar{y}(s)$, $\bar{y} \in L^{\infty}(S, \Sigma, \sigma) \triangleq L^{\infty}(\sigma)$. This $\bar{y}(s)$ is the (second stage) solution of the following "static" formulation of the stochastic programming problem with recourse [16]: minimize over $(x, y(s))$, $x \in R^r$, $y \in L^{\infty}(\sigma)$, the functional

$$f_{10}(x) + \int_S f_{20}(s, x, y(s)) \sigma(ds) \tag{1.12}$$

satisfying conditions

$$x \in C_1, \qquad f_{1j}(x) \leqq 0, \qquad j = 1, \cdots, l_1, \tag{1.13}$$

and almost surely (a.s.)

$$y(s) \in C_2, \qquad f_{2j}(s, x, y(s)) \leqq 0, \qquad j = 1, \cdots, l_2. \tag{1.14}$$

It was shown in [16] that the problems (1.1)-(1.3) and (1.12)-(1.14) are equivalent in the sense that the vector $x \in R^r$ minimizes (1.1)-(1.3) if and only if the pair $(x, y) \in R^r \times L^\infty(\sigma)$ minimizes (1.12)-(1.14), and the optimal values of both problems are equal. Due to this equivalence we can conclude that computational difficulties which arise in numerical solution of stochastic programs with recourse could be in general the same as the difficulties which arise in numerical solution of an extremum problem in a function space with nonlinear operator constraints (constraints (1.14) are just that type). It was pointed out in [15] that "... the static formulation of stochastic programming problems with recourse is computationally more tractable than the dynamic one... and can be solved in some cases by solving a sequence of finite-dimensional discretizations."

In numerical solution of stochastic programming problems with recourse it might be more preferable to use, instead of conditional means $\bar{s}_{in}$, appropriately chosen samples of s. In this purpose we define a convergent quadrature process:

$$\sum_{i=1}^{n} h(s_{in}) m_{in} \to \int_S h(s)\sigma(ds), \quad n \to \infty, \quad \forall h \in C(S). \tag{1.15}$$

It was proved in [21] that for the convergence (1.15) it was necessary and sufficient to have a system of partition $\{\mathscr{A}_n\}$, $\mathscr{A}_n = \{A_{1n}, \cdots, A_{nn}\}$, from Σ with properties

(1) $\sigma(A_{in}) > 0$;
(2) $\bigcup_{i=1}^{n} A_{in} = S$;
(3) $A_{in} \cap A_{jn} = \varnothing, \quad i \neq j$;
(4) $\operatorname{diam} A_{in} \to 0, \quad n \to \infty$;
(5) $\sigma(\partial A_{in}) = 0, \quad i = 1, \cdots, n$;
(6) $\max_{1 \leqq i \leqq n} \dfrac{m_{in}}{\sigma(A_{in})} \to 1, \quad n \to \infty$;
(7) $s_{in} \in A_{in}$,

where ∂A denotes the boundary of a set A and $\operatorname{diam} A = \sup_{s,t \in A} |s - t|$. Our use of the partition $\{\mathscr{A}_n\}$ differs from the use of the partition $\{\mathscr{S}_n\}$ above in the analogous sense as Riemann and Lebesgue integrals differ from each other. Note that the collection of sets $\{\mathscr{A}_n\}$ with properties (1)-(7) generates an algebra $\Sigma_0 \subset \Sigma$ ([21]). Denote the restriction of the measure σ to the algebra Σ_0 by σ_0. Supposing that $S = [0, 1]$ and σ is the Lebesgue measure on $[0, 1]$ then the integrability in the sense of restricted measure σ_0 means simply Riemann integrability.

Suppose that the quadrature process (1.15) converges. Then instead of problem (1.6)-(1.8) we can solve the following mathematical programming problem with a staircase constraint set: minimize over x and y_i, $i = 1, \cdots, n$, the function

$$f_{10}(x) + \sum_{i=1}^{n} f_{20}(s_{in}, x, y_i) m_{in} \tag{1.16}$$

satisfying conditions

$$x \in C_1, \qquad f_{1j}(x) \leqq 0, \qquad j = 1, \cdots, l_1, \tag{1.17}$$

and

$$y_i \in C_2, \qquad f_{2j}(s_{in}, x, y_i) \leqq 0, \qquad j = 1, \cdots, l_2, \quad i = 1, \cdots, n. \tag{1.18}$$

Define by $l_n^\infty(m_n)$ the $v \times n$-dimensional space with norm $\|y_n\|_n = \max_{1 \leqq i \leqq n} |y_{in}|$, where $|\cdot|$ is the Euclidean norm of a v-dimensional vector.

In order to achieve the convergence of $\{\bar{y}_n\}$ to a $\bar{y}(s)$, $\bar{y}_n \in l_n^\infty(m_n)$, $y \in L^\infty(\sigma)$, it is necessary to introduce a system of connection operators $\mathcal{P} = \{p_n\}$, $p_n : L^\infty(\sigma) \to l_n^\infty(m_n)$, $n = 1, 2, \cdots$, between the spaces $l_n^\infty(m_n)$ and $L^\infty(\sigma)$. For the L^P-spaces, $1 \leqq p \leqq \infty$, it is natural to define this system in a piecewise integral form. For any system of connection operators $\mathcal{P} = \{p_n\}$, the following condition should be implemented [20]: for any y

$$\|p_n y\| \to \|y\|, \qquad n \to \infty \tag{1.19}$$

(without (1.19) a discretely converging sequence can have even an infinite number of limits). For any $y \in L^\infty(\sigma)$ the convergence (1.19) takes place [12] and, consequently, makes it possible to solve approximately several problems with a nonlinear operator form $L^\infty(\sigma)$ to $L^\infty(\sigma)$. For example, (1.19) makes it possible to solve approximately continuous programming problems [5], optimal control problems with nonlinear state and control constraints [7] and integral equation [1] in $L^\infty(\sigma)$.

In § 2 we introduce some notions from the theory of discrete convergence necessary for this paper. Since we consider the pair (L^∞, L^1), only discrete and weak* discrete convergences will be used.

In § 3 we deduce conditions that guarantee the convergence of sequence of solutions of discretized problems (1.16)–(1.18) to the solution of the initial problem.

2. Discretization of the problem. Let us introduce some notions from the theory of discrete convergence of mappings (see, e.g., [20], [22]).

Let B and B_n, $n = 1, 2, \cdots$, be Banach spaces with norms $\|\cdot\|$ and $\|\cdot\|_n$, respectively, and let $\mathcal{Q} = \{q_n\}$ be a system of linear connection operators $q_n : B \to B_n$, $n = 1, \cdots$ such that, for every $u \in B$, $\|q_n u\|_n \to \|u\|$ as $n \to \infty$.

DEFINITION 2.1. A sequence $\{u_n\}$ with $u_n \in B_n$ $\mathcal{Q}$-converges (or converges discretely) to $u \in B$ if $\|u_n - q_n u\|_n \to 0$ as $n \to \infty$. We denote this convergence by $u_n \to^{\mathcal{Q}} u$ or simply $u_n \to u$.

DEFINITION 2.2. Connection systems $\mathcal{Q} = \{q_n\}$ and $\mathcal{Q}' = \{q'_n\}$ are called equivalent if for every $u \in B$ $\|q_n u - q'_n u\|_n \to 0$ as $n \to \infty$.

Let B^* and B_n^* be dual spaces to B and B_n, let $\langle u, y \rangle$ denote the value of the linear form $y \in B^*$ at the element $u \in B$ and $\langle u_n, y_n \rangle_n$ denote the value of the linear form $y_n \in B_n^*$ at the element $u_n \in B_n$. Define also a system of linear connection operators $\mathcal{P} = \{p_n\}$, $p_n : B^* \to B_n^*$, $n = 1, \cdots$, such that for every $y \in B^*$ we have

$$\|p_n y\|_n \to \|y\| \quad \text{as } n \to \infty. \tag{2.1}$$

Define (analogously to Definition 2.1) $\mathcal{P}$-convergence of a sequence $\{y_n\}$, $y_n \in B_n^*$, to element $y \in B^*$ and denote this convergence by $y_n \to^{\mathcal{P}} y$ or simply $y_n \to y$.

DEFINITION 2.3. A sequence $\{y_n\}$ with $y_n \in B_n^*$ $\mathcal{Q}$-converges weakly* (or converges weakly* discretely) to $y \in B^*$ if for every discretely converging sequence of elements $u_n \to^{\mathcal{Q}} u$ we have

$$\langle u_n, y_n \rangle_n \to \langle u, y \rangle \quad \text{as } n \to \infty.$$

Denote this convergence by $y_n \dashrightarrow^{\mathcal{Q}} y$ or simply $y_n \dashrightarrow y$.

Denote by $|x_n - x| \to 0$, $n \to \infty$, the convergence in the sense of Euclidean norm.

Define (similarly to the epiconvergence [2], [4]) the discrete epiconvergence of functionals.

Let $(f, \{f_n\})$ be a collection consisting of a functional f and a sequence of functionals $\{f_n\}$. The sequence $\{f_n\}$ is said to $\mathcal{P}$-epiconverge (or epiconverge discretely) to f if for all y we have

$$\liminf_{n\to\infty} f_n(y_n) \geqq f(y) \quad \text{for all } \{y_n\} \ \mathcal{P}\text{-converging to } y; \tag{2.2}$$

there exists $\{y_n\}$ $\mathcal{P}$-converging to y such that

$$\limsup_{n\to\infty} f_n(y_n) \leqq f(y). \tag{2.3}$$

We also need the weak* $\mathcal{Q}$-convergence and a mixture of $\mathcal{P}$- and weak* $\mathcal{Q}$-convergences: if we replace $\mathcal{P}$-convergence of elements by the weak* $\mathcal{Q}$-convergence, the sequence $\{f_n\}$ is said to be $\mathcal{Q}$-weakly* epiconvergent (or epiconvergent weakly* discretely) to f. If we have in (2.2) the $\mathcal{Q}$-weak* convergence and in (2.3) the $\mathcal{P}$-convergence, we say that $\{f_n\}$ epiconverges $\mathcal{Q}\mathcal{P}$-weakly* to f.

Let us be more concrete now about the general notions of discrete convergence to the spaces $L^\infty(\sigma)$ and $L^1(\sigma)$.

Let us restrict the probability measure σ: let the support S of the measure σ be bounded and let

$$\sigma\{s \,|\, |s-t| = \text{const.}\} = 0, \quad \forall t \in S. \tag{A1}$$

For example, the restriction (A1) is fulfilled if the probability measure σ has a density.

For spaces $L^\infty(\sigma)$ and $l_n^\infty(m_n)$, define the system of connection operators $\mathcal{P} = \{p_n\}$, $p_n : L^\infty(\sigma) \to l_n^\infty(m_n)$, $n = 1, \cdots$, in the following way [11]:

$$(p_n y)_{in} = \sigma(A_{in})^{-1} \int_{A_{in}} y(s)\sigma(ds), \qquad i = 1, \cdots, n, \tag{2.4}$$

for an arbitrary fixed collection of sets $\{\mathcal{A}_n\}$, $\mathcal{A}_n = \{A_{1n}, \cdots, A_{nn}\}$ from sigma-field Σ with properties (1)–(7) from the Introduction.

Let $u \in L^1(\sigma)$ and let $\|\cdot\|_1$ denote the norm of an element in $L^1(\sigma)$. Define (analogously to (2.4)) the system $\mathcal{Q}$ of connection operators for spaces $L^1(\sigma)$ and $l_n^1(m_n)$ with a fixed collection of sets $\{\mathcal{B}_n\}$, $\mathcal{B}_n = \{B_{1n}, \cdots, B_{nn}\}$, with properties (1)–(7) (here $l_n^1(m_n)$ is nv-dimensional space with norm

$$\|u_n\|_{n,1} = \sum_{i=1}^{n} |u_{in}| m_{in}.$$

On the set of continuous functions $C(S)$ define (in additon to the system $\mathcal{Q}$) the system of equivalent connection operators $\mathcal{Q}' = \{q_n'\}$:

$$(q_n' z)_{in} = z(s_{in}), \qquad i = 1, \cdots, n, \qquad z \in C(S). \tag{2.4'}$$

Systems of connection operators $\mathcal{Q}$ and $\mathcal{Q}'$ are equivalent according to the Definition 2.2 (see [20, p. 649]). $\mathcal{Q}'$ can be extended onto $L^1(\sigma)$ with preserving linearity of q_n' and property $\|q_n' u\|_{n,1} \to \|u\|_1$ as $n \to \infty$ for every $u \in L^1(\sigma)$.

The system of connection operators $\mathcal{P} = \{p_n\}$ defined by (2.4) satisfies the condition (1.19):

$$\|p_n y\|_n \to \|y\|, \quad n \to \infty, \quad \forall y \in L^\infty(\sigma).$$

The proof of (1.19) for a $y \in L^\infty(\sigma)$ relies on the following two facts:

(1) $\|y\|_p \to \|y\|$, $p \to \infty$ (here $\|y\|_p$ denotes the L^P-norm of $y \in L^\infty(\sigma)$),

(2) $\|\not{p}_n y\|_{n,p} \to \|y\|_p$, $n \to \infty$, for every fixed $p \in [1, \infty)$. The last convergence is equivalent to the following two conditions which can be easily proved:

(2a) $\|\not{p}_n\|_{n,p} \leqq$ const.,

(2a) $\|\not{p}_n y\|_{n,p} \to \|y\|_p$, $n \to \infty$, for any y from $C(S)$ which is dense in $L^p(\sigma)$. The proof of the convergence (2.5) is quite lengthy and is presented in [12].

Remark 2.1. It is easy to see that for the systems of connection operators $\mathcal{P} = \{\not{p}_n\}$ and $\mathcal{Q} = \{\varphi_n\}$ with properties (1)-(7) the compatibility condition [19], $y_n \to^{\mathcal{P}} y$, $u_n \to^{\mathcal{Q}} u \Rightarrow \langle u_n, y_n \rangle_n \to \langle u, y \rangle$ is fulfilled (here $y \in L^\infty(\sigma)$, $y_n \in l_n^\infty(m_n)$, $u \in L^1(\sigma)$, $u_n \in l_n^1(m_n)$).

3. Conditions for convergence of discrete approximations. Let us introduce some notations and assumptions necessary for the discrete approximation of the problem (1.12)-(1.14) by the sequence of problems (1.16)-(1.18). Denote by G and G_n the (nonempty) sets of admissible solutions of problems (1.12)-(1.14) and (1.16)-(1.18), respectively, and by $G_2(s, x)$ the following set:

$$G_2(s, x) = \{y \mid f_{2j}(s, x, y) \leqq 0, j = 1, \cdots, l_2\}.$$

Suppose that

(A2) the functions f_{1j} on R^r, $j = 0, 1, \cdots, l_1$, are convex and differentiable;

(A3) the functions $f_{2j}(s, \cdot, \cdot)$ on $R^r \times R^v$, $j = 0, 1, \cdots, l_2$, are convex and differentiable, for each (x, y) on $R^r \times R^v$ the functions $f_{2j}(\cdot, x, y)$, $j = 0, 1, \cdots, l_2$, are bounded and measurable on S. Moreover, to each bounded set $B \subset R^r \times R^v$ there corresponds a bounded and Σ_0-measurable function $\alpha: S \to R$ and a constant $\beta \in R$ such that $|f_{20}(s, x, y)| \leqq \alpha(s)$ for all $(x, y) \in B$, $|f_{2j}(s, x, y)| \leqq \beta$ for all $(x, y) \in B$, $j = 1, \cdots, l_2$;

(A4) the functions $f'_{2jx}(s, \cdot, \cdot)$, $f'_{2jy}(s, \cdot, \cdot)$ on $R^r \times R^v$, $j = 0, 1, \cdots, l_2$, are continuous, for each (x, y) on $R^r \times R^v$ the functions $f'_{2jy}(\cdot, x, y)$ are bounded and Σ_0-measurable on S. Moreover, to each bounded set $B \subset R^r = R^v$ there correspond bounded and Σ_0-measurable functions $\gamma_j: S \to R$, $\delta_j: S \to R$, such that

$$|f'_{2jx}(s, x, y)| \leqq \gamma_j(s) \quad \text{for all } (x, y) \in B,$$

$$|f'_{2jy}(s, x, y)| \leqq \delta_j(s) \quad \text{for all } (x, y) \in B.$$

Remark 3.1. In order to guarantee the convergence of discrete approximations, more stringent conditions (A3), (A4) compared with [18] are needed. The same is valid about the paper [6].

(A5) The sets $C_1 \subset R^r$ and $C_2 \subset R^v$ are bounded closed and convex with int $C_2 \neq \emptyset$;

(A6) there exists an element $(\tilde{x}, \tilde{y}(s))$ such that for some $\varepsilon > 0$ the constraints (1.2) and (almost surely) (1.3) can be satisfied with $-\varepsilon$ in place of 0 (strict feasibility [17]);

(A7) the support S of the measure σ is bounded; let

$$K_1 = \{x \mid x \in C_1, f_{1j}(x) \leqq 0, j = 1, \cdots, l_1\}.$$

(A8) For all $x \in K_1$ there is bounded region $D \subset R^v$ with $G_2(s, x) \cap D \neq \emptyset$ for all $s \in S$ (relatively complete recourse [17]).

With assumptions (A2), (A3), and (A5), the optimal recourse problem (1.1)-(1.3) is well defined (see, e.g., [18]).

Since the existence conditions do not depend on the properties of the measure σ (discrete measures are not excluded) under the conditions above the problems (1.16)-(1.18) are well defined.

Define together with the connection system $\mathcal{P}$ the following system of piecewise constant restoration operators $\mathcal{R}=\{\imath_n\}$, $\imath_n: l_n^\infty(m_n)\to L^\infty(\sigma)$, of the form

$$(\imath_n y_n)(s)=y_{in} \quad \text{as} \quad s\in A_{in} \tag{3.1}$$

where the sets $\{A_{in}\}_{i=1}^n$ are taken from the system $\mathcal{P}$ with properties (1)-(7) (see (2.4)).

PROPOSITION 3.1. *Let $f_{1j}, j=1,\cdots,l_1$, satisfy the condition* (A2), $f_{2j}, j=1,\cdots,l_2$, *conditions* (A3), (A4), *sets C_1, C_2 the condition* (A5). *If the convergence* (1.15) *holds, then the constraint set G contains all weak* discrete limits of sequences $\{(x_k, y_k)\}$, $(x_k, y_k)\in G_{n_k}$, where $\{G_{n_k}\}$ are arbitrary subsequences of $\{G_n\}$.*

Proof. Let $|x_n - x|\to 0$ and $y_n \to y$ as $n\to\infty$. Consider first constraints (1.18). Let $f_{2j}(s_{in}, x_n, y_{in})\leqq 0$ for all $j=1,\cdots,l_2$ and all $i=1,\cdots,n$, $n=1,2,\cdots$. Let us explain the idea of the proof. On the contrary suppose that there exists a set $D\in\Sigma$ with a positive measure $\sigma(D)>0$, and an index $j\in\{1,\cdots,l_2\}$ such that $f_{2j}(s, x, y(s))\geqq\delta>0$ for all $s\in D$. Then $\delta\sigma(D)\leqq\int_S \chi_D(s) f_{2j}(s, x, y(s))\sigma(ds)$ where

$$\chi_D(s)=\begin{cases}1, & s\in D,\\ 0, & s\notin D.\end{cases}$$

Then the function $\chi_D(s) f_{2j}(s, x, y(s))$ is approximated (as an element of $L^1(\sigma)$) by the function $z(s) f_{2j}(s, x, y_c(s))$, where z, y_c are continuous functions. This enables us to estimate the difference

$$\sum_{i=1}^n |(\wp_n(\chi_D(s) f_{2j}(s, x, y(s)))_{in} - (\wp_n\chi_D(s))_{in} f_{2j}(s_{in}, x, (\wp_n y(s))_{in})| m_{in}. \tag{3.2}$$

If the difference (3.2) is sufficiently small we have the contradiction

$$\begin{aligned} 0\leqq\delta\sigma(D) &\leqq \int_S \chi_D(s) f_{2j}(s, x, y(s))\sigma(ds)\\ &\leqq \sum_{i=1}^n (\wp_n\chi_D(s))_{in} f_{2j}(s_{in}, x_n, y_{in}) m_{in} + \frac{1}{2}\delta\sigma(D) \qquad (3.3)\\ &\leqq \frac{1}{2}\delta\sigma(D) \quad \text{for } n \text{ sufficiently large.}\end{aligned}$$

After these introductory remarks let us now prove Proposition 3.1.

Since $\chi_D(\cdot) f_{2j}(\cdot, x, y(\cdot))\in L^1(\sigma)$ and (definition of $\mathcal{P}$-convergence)

$$\sum_{i=1}^n (\wp_n(\chi_D(s) f_{2j}(s, x, y(s)))_{in} m_{in} \to \int_S \chi_D(s) f_{2j}(s, x, y(s))\sigma(ds), \quad n\to\infty,$$

we have for $n\geqq n_1$

$$\left|\sum_{i=1}^n (\wp_n(\chi_D(s) f_{2j}(s, x, y(s)))_{in} m_{in} - \int_S \chi_D(s) f_{2j}(s, x, y(s))\sigma(ds)\right| \leqq \delta/8\sigma(D).$$

For brevity denote $f_{2j}(x, y)=f_{2j}(s, x, y(s))$ and $(\wp_n\chi_D)_{in} f_{2jn}(x, (\wp_n y)_n)_{in} = (\wp_n\chi_D(s))_{in} f_{2j}(s_{in}, x, (\wp_n y)_{in})$. Consider the difference (3.2):

$$\begin{aligned} &\|(\wp_n(\chi_D f_{2j}(x, y)))_n - (\wp_n\chi_D)_n f_{2jn}(x, (p_n y)_n)\|_n\\ &\leqq \sum_{i=1}^n |(\wp_n(\chi_D f_{2j}(x, y)))_n - (\wp_n(z f_{2j}(x, y_c)))_n| m_{in}\\ &\quad + \sum_{i=1}^n (\wp_n(z f_{2j}(x, y_c)))_n - (\wp_n\chi_D)_n f_{2jn}(x, (\wp_n y)_n)| m_{in}\end{aligned}$$

for some z, $y_c \in C(S)$. Take continuous $z(s)$, $0 \leqq z(s) \leqq 1$, $s \in S$, and $y_c(s)$ such that

$$\int_S |\chi_D(s) - z(s)|\sigma(ds) < \delta \Big/ \left(64 \sup_{s \in S} |f_{2j}(s, x, y_c(s))|\right)$$

and

$$\int_S |f_{2j}(s, x, y(s)) - f_{2j}(s, x, y_c(s))|\sigma(ds) < \delta/64$$

(for any $\nu > 0$ one can choose in $L^1(\sigma)$ such a continuous $z(s)$ that $\int_S |\chi_D(s) - z(s)|\sigma(ds) < \upsilon$). Note that $\sup_{s \in S} |f_{2j}(s, x, y_c(s))|$ is finite since the function $g(s) = |f_{2j}(s, x, y_c(s))|$ is bounded and Σ_0-measurable.

Then for $n \geqq n_2$

$$\sum_{i=1}^{n} |(\not{p}_n(\chi_D f_{2j}(x, y)))_n - (\not{p}_n(z f_{2j}(x, y_c)))_n| m_{in}$$

$$= \sum_{i=1}^{n} |(\not{p}_n(\chi_D f_{2j}(x, y) - z f_{2j}(x, y_c))_n| m_{in}$$

$$\leqq \int_S |\chi_D(s) f_{2j}(s, x, y(s)) - z(s) f_{2j}(s, x, y_c(s))|\sigma(ds)$$

$$+ (\delta/32)\sigma(D) \leqq (\delta/16)\sigma(D).$$

Consider now the difference:

$$\sum_{i=1}^{n} |(\not{p}_n(z f_{2j}(x, y_c)))_{in} - (\not{p}_n \chi_D)_{in} f_{2jn}(x, (\not{p}_n y)_n)| m_{in}$$

$$\leqq \sum_{i=1}^{n} |(\not{p}_n(z f_{2j}(x, y_c)))_{in} - (\not{p}_n z)_{in} f_{2jn}(x, (\not{p}_n y_c)_n))| m_{in}$$

$$+ \sum_{i=1}^{n} |f_{2jn}(x, (\not{p}_n y)_n) - f_{2jn}(x, (\not{p}_n y_c)_n)| m_{in}$$

$$+ \max_{1 \leqq i \leqq n} |f_{2jn}(x, (\not{p}_n y_c)_n)| \sum_{i=1}^{n} |(\not{p}_n(\chi_D - z))_{in}| m_{in}$$

$$\leqq (\delta/16)\sigma(D)$$

as $n \geqq n_3$ (due to the boundedness of $\max_{1 \leqq i \leqq n} |f_{2jn}(x, (\not{p}_n y_c)_n)|$ and the inequality

$$\sum_{i=1}^{n} |(\not{p}_n(\chi_D - z))_{in}| m_{in} \leqq \int |\chi_D(s) - z(s)|\sigma(ds)$$

$$+ \delta \Big/ \left(48 \max_{1 \leqq i \leqq n} |f_{2jn}(x, (\not{p}_n y_c)_n)|\right).$$

Taking $n_4 = \max\{n_2, n_3\}$ we obtain for $n \geqq n_4$ that

$$\|(\not{p}_n(\chi_D f_{2j}(x, y)))_n - (\not{p}_n \chi_D)_n f_{2jn}(x, (\not{p}_n y)_n)\|_n < (\delta/8)\sigma(D).$$

Since $\not\pi_n y \to y$, $n = 1, \cdots, y_{n_k} \dashrightarrow y$, $|x_{n_k} - x| \to 0$, $k = 1, \cdots$, and sequences $\{(\not\pi_n \chi_D)_{in} f'_{2jx}(s_{in}, x, (\not\pi_n y)_{in})\}$, $\{(\not\pi_n \chi_D)_{in} f'_{2jy}(s_{in}, x, (\not\pi_n y)_{in})\}$ converge (analogously to the convergence (3.2)) discretely, then for $n \geqq n_5$ we have

$$\sum_{i=1}^{n} (\not\pi_n \chi_D)_{in} f_{2j}(s_{in}, x, (\not\pi_n y)_{in}) m_{in}$$

$$\leqq \sum_{i=1}^{n} (\not\pi_n \chi_D)_{in} f_{2j}(s_{in}, x_n, y_{in}) m_{in} + (\delta/4)\sigma(D)$$

$$\leqq (\delta/4)\sigma(D)$$

((x_n, y_n) is admissible for (1.16)-(1.18)). Then for $n \geqq n_0 = \max\{n_1, n_4, n_5\}$ we reached the contradiction: $\delta\sigma(D) \leqq \frac{1}{2}\delta\sigma(D)$. So we can conclude that for the limit point $(x, y(s))$ we have $f_{2j}(s, x, y(s)) \leqq 0$ for almost all $s \in S$ and all $j = 1, \cdots, l_2$.

Since $(\imath_n y_n)(s) \in C_2$ for almost all $s \in S$ then due to the closedness of C_2, the limit point $y(s) \in C_2$ for almost all $s \in S$. Conditions (A5) to C_1 and (A2) to $f_{1j}, j = 1, \cdots, l_1$, guarantee that $(x, y(s))$ satisfies all constraints (1.17), (1.18). □

Remark 3.2. Let $C_2 = \{y \mid Dy \geqq d\}$ be a bounded polyhedron with $D - a$ $(k \times v)$-matrix and $d - k$-dimensional vector. Then clearly $(\imath_n y_n)(s) \in C_2$ for almost all $s \in S$ and all $n = 1, 2, \cdots$.

Remark 3.3. The Σ_0-measurability assumption enables us to replace the functions f_{2j}, $j = 0, 1, \cdots, l_2$, and its derivatives in approximation process by their values in discretization points s_{in}. Assuming only Σ-measurability we must use conditional means $\bar{s}_{in}$ instead of s_{in} (see the Introduction). If the functions f_{2j}, $j = 0, 1, \cdots, l_2$, are only Σ-measurable then changing their values on a set of measure zero it is possible that sums in (3.3) are equal to zero for all $n = 1, 2, \cdots$, but the value of the integral does not change. For the same reasons the discrete approximation scheme proposed in [14] may diverge.

Remark 3.4. To make the proof of Proposition 3.1 simpler we could suppose the weak* continuity of $f_{2j}, j = 1, \cdots, l_2$, relative to y. Unfortunately, such a superposition operator is weakly continuous in L^P-spaces, $1 \leqq p < \infty$ (weakly* continuous in L^∞), if and only if it is linear.

Define function F_0 and F_{0n}:

$$F_0 = f_{10}(x) + \int_S f_{20}(s, x, y(s))\sigma(ds),$$

$$F_{0n} = f_{10}(x) + \sum_{i=1}^{n} f_{20}(s_{in}, x, y_{in}) m_{in}.$$

PROPOSITION 3.2. *Let f_{10} satisfy the condition* (A2), *f_{20} conditions* (A2), (A4). *Let the convergence* (1.15) *hold. Then the sequence $\{F_{0n}\}$ $2\mathcal{P}$-weakly* epiconverges to F_0.*

Proof. Let us show that the collection $(F_0, \{F_{0n}\})$ satisfies (2.3). Let $|x_n - x| \to 0$ and $y_n \xrightarrow{\mathcal{P}} y$ as $n \to \infty$. Then

$$f_{10}(x_n) + \sum_{i=1}^{n} f_{20}(s_{in}, x_n, y_{in}) m_{in} - f_{10}(x) - \int f_{20}(s, x, y(s))\sigma(ds)$$

$$\leqq f_{10}(x_n) - f_{10}(x) + |x_n - x| \sum_{i=1}^{n} |f'_{20x}(s_{in}, x_n, y_{in})| m_{in}$$

$$+ \max_{1 \leqq i \leqq n} |y_{in} - (\not\pi_n y)_{in}| \sum_{i=1}^{n} |f'_{20y}(s_{in}, x_n, y_{in})| m_{in} + R_{0n}(x, y)$$

where

$$R_{0n}(x, y) = \left| \sum_{i=1}^{n} f_{20}(s_{in}, x_n, (\not\pi_n y)_{in}) m_{in} - \int f_{20}(s, x, y(s))\sigma(ds) \right|.$$

Since $|x_n - x| \to 0$ and $\not\pi_n y \to^{\mathscr{P}} y$, $n \to \infty$, by the definition, then

$$\left| \sum_{i=1}^{n} (\not\pi_n f_{20}(x, y))_{in} m_{in} - \int f_{20}(s, x, y(s))\sigma(ds) \right| \to 0$$

and

$$\sum_{i=1}^{n} |(\not\pi_n f_{20}(x, y)_{in} - f_{20n}(x_n, (\not\pi_n y)_{in})| m_{in} \to 0, \quad n \to \infty$$

(as in the proof of (3.2) in Proposition 3.1). Consequently, $R_{0n}(x, y) \to 0$ as $n \to \infty$. Due to the convergences $|x_n - x| \to 0$ and $y_n \to^{\mathscr{P}} y$, the collection $(F_0, \{F_{0n}\})$ satisfies (2.3) as $n \to \infty$.

Consider the convergence (2.2). Since a bounded sequence in $L^{\infty}(\sigma)$ is weakly* compact, we have to use the $\mathscr{Q}$-weak* convergence in (2.2). Let $|x_n - x| \to 0$ and $\langle u_n, y_n \rangle_n \to \langle u, y \rangle$ as $n \to \infty$ for every $u_n \to^{\mathscr{Q}} u$, $u_n \in l_n^1(m_n)$, $u \in L^1(\sigma)$. Due to the condition (A4) $f'_{0y} \triangleq f'_{20y}(\cdot, x, y(\cdot)) \in L^1(\sigma)$ also $f'_{0ny} \triangleq f'_{20y}(s_{in}, x, (\not\pi_n y)_{in}) \in l_n^1(m_n))$. Then, analogously to convergence $R_{0n}(x, y) \to 0$, $n \to \infty$, we can show that the sequence $\{f'_{0ny}\}$ converges discretely to $f'_{0y} \in L^1(\sigma)$.

Consider now the difference

$$f_{10}(x) + \int f_{20}(s, x, y(s))\sigma(ds) - f_{10}(x_n) - \sum_{i=1}^{n} f_{20}(s_{in}, x_n, y_{in}) m_{in}$$

$$\leqq f_{10}(x) - f_{10}(x_n) + \sum_{i=1}^{n} (f'_{20x}(s_{in}, x, (\not\pi_n y)_{in}), x - x_n) m_{in}$$

$$+ \sum_{i=1}^{n} (f'_{20y}(s_{in}, x, (\not\pi_n y)_{in}), (\not\pi_n y)_{in} - y_{in}) m_{in} + R_{0n}(x, y).$$

In the last sum all components tend to zero as $n \to \infty$ (by the definition $\not\pi_n y \to^{\mathscr{P}} y$ and by the assumption $y_n \dashrightarrow^{\mathscr{Q}} y$). Hence, the collection $(F_0, \{F_{0n}\})$ satisfies (2.2) as $|x_n - x| \to 0$, $y_n \dashrightarrow^{\mathscr{Q}} y$, $n \to \infty$. □

Let us now formulate and prove the main result of the paper.

Denote by F^* the optimal value of the problem (1.12)-(1.14), by $(\bar{x}_n, \bar{y}_n)$ and F_n^* the solution and optimal value of the problem (1.16)-(1.18), respectively.

THEOREM 3.1. *Let the conditions* (A1)-(A8) *be fulfilled. Let the convergence* (1.15) *hold. Then*

$$\lim_{n \to \infty} F_n^* = F^*$$

and all $\mathscr{Q}$-weak limit points $(\bar{x}, \bar{y})$ of the sequence $\{(\bar{x}_n, \bar{y}_n)\}$ of solutions of the problems* (1.16)-(1.18) *solve the problem* (1.12)-(1.14).

Proof. By admissibility of limit point $(\bar{x}, \bar{y})$ and by Proposition 3.2 we have

$$F^* \leqq f_{10}(\bar{x}) + \int f_{20}(s, \bar{x}, \bar{y}(s))\sigma(ds)$$

$$\leqq \liminf_{n \to \infty} \left\{ f_{10}(\bar{x}_n) + \sum_{i=1}^{n} f_{20}(s_{in}, \bar{x}_n, \bar{y}_{in}) m_{in} \right\}$$

$$= \liminf_{n \to \infty} F_n^*.$$

Let us show that the opposite inequality holds.

Due to the strict feasibility condition (A6) there exists an admissible element $(x_\lambda, y_\lambda(s))$, $x_\lambda = \lambda\tilde{x} + (1-\lambda)\bar{x}$, $y_\lambda = \lambda\tilde{y} + (1-\lambda)\bar{y}$, $0 \leqq \lambda \leqq 1$, such that

(1) $|f_{10}(x_\lambda) + \int f_{20}(s, x_\lambda, y_\lambda(s))\sigma(ds) - f_{10}(\bar{x}) - \int f_{20}(s, \bar{x}, \bar{y}(s))\sigma(ds)| < \varepsilon/2$,

(2) for $n \geqq n_1$ we have $(x_\lambda, \not{p}_n y_\lambda) \in G_n$.

Indeed, consider inequality (1):

$$\left| f_{10}(x_\lambda) + \int f_{20}(s, x_\lambda, y_\lambda(s))\sigma(ds) - f_{10}(\bar{x}) - \int f_{20}(s, \bar{x}, \bar{y}(s))\sigma(ds) \right|$$

$$\leqq \lambda \left| f_{10}(\tilde{x}) + \int f_{20}(s, \tilde{x}, \tilde{y}(s))\sigma(ds) \right.$$

$$\left. - f_{10}(\bar{x}) - \int f_{20}(s, \bar{x}, \bar{y}(s))\sigma(ds) \right|.$$

Take λ so little that

$$\left| f_{10}(x_\lambda) + \int f_{20}(s, x_\lambda, y_\lambda(s))\sigma(ds) - f_{10}(\bar{x}) - \int f_{20}(s, \bar{x}, \bar{y}(s))\sigma(ds) \right| \leqq \varepsilon/2.$$

Now $\lambda = \lambda(\varepsilon)$ is fixed. In order to guarantee the inclusion (2), $(x_\lambda, \not{p}_n y_\lambda) \in G_n$ for all $n \geqq n_1$, n_1-sufficiently large, note that the sequence $\{f_{2j}(s_{in}, x, (\not{p}_n y)_{in})\}$ $\mathcal{P}$-converges to $f_{2j}(s, x, y(s))$, $j = 1, \cdots, l_2$, i.e.,

$$\max_{1 \leqq i \leqq n} |f_{2j}(s_{in}, x, (\not{p}_n y)_{in}) - (\not{p}_n f_{2j}(s, x, y(s)))_{in}| \to 0, \quad n \to \infty. \tag{3.4}$$

This convergence is guaranteed by properties of function f_{2j} (continuity in (x, y) and Σ_0-measurability in (s) and by the definition of $\mathcal{P}$-convergence of the sequence $\{\not{p}_n y\}$ to y. Consequently,

$$\begin{aligned} f_{2j}(s_{in}, x_\lambda, (\not{p}_n y_\lambda)_{in}) \leqq{}& \lambda(f_{2j}(s_{in}, \tilde{x}, (\not{p}_n \tilde{y})_{in}) \\ & - (\not{p}_n f_{2j}(s, \tilde{x}, \tilde{y}(s)))_{in}) + \lambda(\not{p}_n f_{2j}(s, \tilde{x}, \tilde{y}(s)))_{in} \\ & + (1-\lambda)(f_{2j}(s_{in}, \bar{x}, (\not{p}_n \bar{y})_{in}) - (\not{p}_n f_{2j}(s, \bar{x}, \bar{y}(s)))_{in}) \\ & + (1-\lambda)(\not{p}_n f_{2j}(s, \bar{x}, \bar{y}(s)))_{in}. \end{aligned}$$

Since $(\not{p}_n f_{2j}(s, \bar{x}, \bar{y}(s)))_{in} \leqq 0$, $j = 1, \cdots, l_2$, $i = 1, \cdots, n$, $(\not{p}_n f_{2j}(s, \tilde{x}, \tilde{y}(s)))_{in} \leqq -\varepsilon < 0$, $j = 1, \cdots, l_2$, $i = 1, \cdots, n$, and due to $\mathcal{P}$-convergences (3.4) for arguments $(\tilde{x}, \tilde{y}(s))$ and $(\bar{x}, \bar{y}(s))$ we can conclude that there exists a sufficiently large index n_1 such that $(x_\lambda, \not{p}_n y_\lambda) \in G_n$ for all $n \geqq n_1$.

Then from the admissibility of the point $(x_\lambda, \not{p}_n y)$ we obtain the inequality

$$\begin{aligned} & f_{10}(\bar{x}_n) + \sum_{i=1}^{n} f_{20}(s_{in}, \bar{x}_n, \bar{y}_{in}) m_{in} - f_{10}(\bar{x}) - \int f_{20}(s, \bar{x}, \bar{y}(s))\sigma(ds) \\ & \leqq \left\{ f_{10}(x_\lambda) + \sum_{i=1}^{n} f_{20}(s_{in}, x_\lambda, (\not{p}_n y_\lambda)_{in}) m_{in} \right\} \\ & \quad - \left\{ f_{10}(x_\lambda) + \int f_{20}(s, x_\lambda, y_\lambda(s))\sigma(ds) \right\} \\ & \quad + \left| f_{10}(x_\lambda) + \int f_{20}(s, x_\lambda, y_\lambda(s))\sigma(ds) - f_{10}(\bar{x}) - \int f_{20}(s, \bar{x}, \bar{y}(s))\sigma(ds) \right|. \end{aligned}$$

Due to the $\mathcal{P}$-convergence (2.3) of the collection $(F_0, \{F_{0n}\})$ and continuity of f_{10} the first difference in the last sum does not exceed $\varepsilon/2$ as $n \geqq n_2$.

Then for $n \geqq \max\{n_1, n_2\}$ we have

$$f_{10}(\bar{x}_n) + \sum_{i=1}^{n} f_{20}(s_{in}, \bar{x}_n, \bar{y}_{in}) m_{in} \leqq f_{10}(\bar{x}) + \int f_{20}(s, \bar{x}, \bar{y}(s)) \sigma(ds) + \varepsilon.$$

Consequently, $\lim_{n\to\infty} F_n^* = F^*$.

The remaining part of the theorem follows from the $\mathcal{Q}$-weak* convergence of $\{(\bar{x}_{n_k}, \bar{y}_{n_k})\}$, $k = 1, \cdots$, and the admissibility of its limit point $(\bar{x}, \bar{y}(s))$.

Example. Let us illustrate the idea of discrete approximation with a simple example. Consider the Example 2′ (with relatively complete recourse) from [17]: find $x \in R^1$ and $y \in L^\infty[0, 1]$ such that $x \geqq 1$, $y(s) \geqq 0$ and $y(s) - x + s \leqq 0$ for almost all s minimizing the expression

$$2x - \int_0^1 y(s)\, ds.$$

Here $\bar{x} = 1$, $\bar{y}(s) = 1 - s$, $F^* = \frac{3}{2}$.

Consider now the discretized problem: Find $x \in R^1$ and $y_{in} \in R^1$, $i = 1, \cdots, n$, such that $x \geqq 1$, $y_{in} \geqq 0$ and $y_{in} - x + i/n \leqq 0$, $i = 1, \cdots, n$, minimizing the expression

$$2x - \frac{1}{n} \sum_{i=1}^{n} y_{in}.$$

Clearly $\bar{x}_n = 1$, $\bar{y}_{in} = 1 - i/n$, $F_n^* = 3/2 + 1/2n$. Hence, $F_n^* \to F^*$, $\bar{x}_n = \bar{x}$ and $\bar{y}_n \to \bar{y}$ discretely.

4. Concluding remarks. The aim of this paper was to develop an approximation scheme for stochastic convex programs with relatively complete recourse. The problem in the space $R^r \times L^\infty(\sigma)$ was replaced by a mathematical programming problem in a finite-dimensional space. This replacement is justified by the notion of discrete convergence of mappings which describes the convergence of elements in different spaces by a system of connection operators.

Unfortunately, we are not able to solve approximately the general stochastic convex program with recourse. In general the "singular multipliers" [17] in Kuhn-Tucker conditions result from the presence of induced constraints and we have to consider the problem (1.12)-(1.14) in L^∞ in pairing with $(L^\infty)^* \asymp ba(S, \Sigma, \sigma)$—the space of bounded additive set functions which are absolutely continuous with respect to σ. Since for an element $\mu \in ba(S, \Sigma, \sigma)$ the Radon-Nikodym Theorem is formulated under extremely strong conditions [13] (without the Radon-Nikodym Theorem we are not able to discretize the set function μ from $ba(S, \Sigma, \sigma)$), it is quite difficult to define the discrete convergence in $(L^\infty)^*$.

Acknowledgments. The author expresses his gratitude to the referees for the constructive criticism concerning stochastic programming and discrete approximation problems and to the editor for the encouraging attitude.

REFERENCES

[1] K. E. ATKINSON AND F. A. POTRA, *Projection method for nonlinear integral equations*, SIAM J. Numer. Anal., 24 (1987), pp. 1352-1373.

[2] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, Applicable Mathematics Series, London, 1984.

[3] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[4] R. J. BIRGE AND R. J-B. WETS, *Designing approximation schemes for stochastic optimization problems, in particular for stochastic programs with recourse*, Math. Programming Stud., 27 (1986), pp. 54–102.

[5] R. N. BUIE AND J. ABRAHAM, *Some remarks concerning duality for continuous-time programming problems*, J. Math. Anal. Appl., 114 (1986), pp. 468–489.

[6] S. D. FLÅM, *Nonanticipativity in stochastic programming*, J. Optim. Theory Appl., 46 (1985), pp. 23–30.

[7] W. W. HAGER AND S. K. MITTER, *Lagrange duality theory for convex control problems*, SIAM J. Control Optim., 14 (1976), pp. 843–856.

[8] P. KALL, *On approximation and stability in stochastic programming*, in Parametric Optimization and Related Topics, J. Guddat, ed., Akademie-Verlag, Berlin, 1987, pp. 387–407.

[9] P. KALL, K. FRAUENDORFER, AND A. RUSZCYNSKI, *Approximation techniques in stochastic programming*, Report, Institut für Operations Research der Universität Zürich, Zürich, Switzerland, October 1984; in Numerical Techniques and Stochastic Optimization, Y. Ermoliev, R. Wets, eds., Springer-Verlag, Berlin, 1988, to appear.

[10] P. KALL AND D. STOYAN, *Solving stochastic programming problems with recourse including error bounds*, Math. Operationsforsch. Statist., Ser. Optimization, 13 (1982), pp. 431–447.

[11] R. LEPP, *Discrete approximation of linear two-stage stochastic programming problem*, Numer. Funct. Anal. Optim., 9 (1987), pp. 19–33.

[12] ———, *Discrete approximation conditions for the space of essentially bounded functions*, Proc. Acad. Sci. Estonian SSR. Phys.-Math., 37 (1988), pp. 204–208 (in Russian).

[13] H. B. MAYNARD, *A Radon-Nikodym theorem for finitely additive bounded measures*, Pacific J. Math., 83 (1979), pp. 401–413.

[14] P. OLSEN, *Discretization of multistage stochastic programming problems*, Math. Programming Stud., 6 (1976), pp. 111–124.

[15] ———, *Multistage stochastic programming with recourse as mathematical programming in an L_p space*, SIAM J. Control Optim., 14 (1976), pp. 528–537.

[16] R. T. ROCKAFELLAR AND R. J-B. WETS, *Stochastic convex programming: Basic duality*, Pacific J. Math., 62 (1976), pp. 173–195.

[17] ———, *Stochastic convex programming: relatively complete recourse and induced feasibility*, SIAM J. Control Optim., 14 (1976), pp. 574–589.

[18] ———, *The optimal recourse problem in discrete time: L^1-multipliers for inequality constraints*, SIAM J. Control Optim., 16 (1978), pp. 16–36.

[19] F. STUMMEL, *Stability and discrete convergence of differential mappings*, Rev. Roumaine Math. Pures Appl., 21 (1976), pp. 63–96.

[20] G. VAINIKKO, *Approximative methods for nonlinear equations (two approaches to the convergence problem)*, Nonlinear Anal., 2 (1978), pp. 647–687.

[21] ———, *On convergence of quadrature formulae method for integral equations with discontinuous kernels*, Sibirsk. Mat. Zh., 12 (1971), pp. 40–53 (in Russian).

[22] V. V. VASIN, *Discrete approximation and stability in extremal problems*, Zh. Vychisl. Mat. i Mat. Fiz., 22 (1982), pp. 824–839 (in Russian).

[23] R. J-B. WETS, *Stochastic programming: solution techniques and approximation schemes*, in Mathematical Programming Bonn 1982: The State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 566–603.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 395-401, March 1990

008

ON A THEOREM OF ARROW, BARANKIN, AND BLACKWELL*

MARTIN PETSCHKE†

Abstract. A well-known theorem of Arrow, Barankin, and Blackwell states that if $\mathbf{R}^n$ is equipped with the natural ordering, then for every compact convex subset S of $\mathbf{R}^n$ the set of properly minimal elements of S is dense in the set of minimal elements of S.

In this note a result of Jahn is used to show a generalization of the density theorem of Arrow, Barankin, and Blackwell. It will be shown that this theorem holds in a real normed space that is partially ordered by a convex cone with a closed bounded base.

Key words. vector optimization, convex cones

AMS(MOS) subject classifications. 49A27, 46A40, 52A07

1. Introduction. A well-known theorem of Arrow, Barankin, and Blackwell [1] states that if $\mathbf{R}^n$ is equipped with the natural ordering, then for every compact convex subset S of $\mathbf{R}^n$ the set of properly minimal elements of S is dense in the set of minimal elements of S. It was shown by Bitran and Magnanti in [2, Cor. 3.1] that this result remains valid if $\mathbf{R}^n$ is partially ordered by an arbitrary closed convex pointed cone. The Arrow-Barankin-Blackwell Theorem was extended by Borwein [3, Thm. 2] to real normed spaces partially ordered by a convex cone with weakly compact base. In [9, Satz 1] Salz proved the result of Arrow-Barankin-Blackwell in real normed spaces with base norms.

Recently, Jahn [5] has shown that the Arrow-Barankin-Blackwell Theorem remains true in real normed spaces, partially ordered by a Bishop-Phelps cone. What the previously mentioned results have in common is that the ordering cones have closed bounded bases. So we might conjecture that the Arrow-Barankin-Blackwell Theorem is valid for real normed spaces partially ordered by a convex cone with a closed bounded base.

In this note we will answer this conjecture in the affirmative (Theorem 4.1, Corollary 4.2). The idea of our proof is to extend the scope of Jahn's result by means of a characterization lemma concerning Bishop-Phelps cones.

Another problem that arises in this context is the following question. Jahn has shown that the Arrow-Barankin-Blackwell result is valid in L_1 and l_1 spaces, and Borwein has proven it in the case of reflexive spaces. What can be stated for a finite product of some L_1-spaces and some reflexive spaces? In general this product space is neither of L_1 type nor is it reflexive. We will cover this question in an example of a cooperative n player game in § 5.

The result of Corollary 4.2 has an interesting consequence in vector optimization, since it shows that a numerical solution of a convex problem can be computed by solving appropriate scalarized problems.

2. Minimality, proper minimality and Bishop–Phelps cones. Before we present the precise statements, let us recall some basic notions and definitions. In this section let E denote a locally convex Hausdorff space partially ordered by a convex cone C. A subset B of C is called a *base* of C, if B is convex and if each $x \in C\backslash\{0\}$ has a unique

* Received by the editors October 21, 1987; accepted for publication (in revised form) May 2, 1989.

† Technische Hochschule Darmstadt, Fachbereich Mathematik, AG5-Funktionalanalysis, Schloßgartenstrasse 7, D 6100 Darmstadt, Federal Republic of Germany.

representation $x = \lambda b$ for some $\lambda > 0$ and some $b \in B$. Let us note that if C is a nontrivial cone with basis B then we have $0 \notin B$. We call C *pointed*, if $C \cap -C = \{0\}$. Throughout this paper let E^* denote the topological dual space of E. The *dual cone* C^* of C is defined as

$$C^* := \{l \in E^* : l(x) \geqq 0 \ (x \in C)\}.$$

An important subset of C^* is the *quasi-interior* of C^*, which we denote by $C^{\#}$:

$$C^{\#} := \{l \in E^* : l(x) > 0 \ (x \in C \setminus \{0\})\}.$$

If C has a closed bounded base then C^* has nonempty interior in the strong topology of E^* [7, 3.8.4]. All those interior points are in $C^{\#}$. Now let S be a nonempty subset of E. The convex hull of S will be denoted by conv. (S). We call $\bar{x} \in S$ a *minimal* element of S (with respect to C), if

$$S \cap (\bar{x} - C) = \{\bar{x}\}.$$

An element $\bar{x} \in S$ is called a *properly minimal* element of S (with respect to C) if there is some $l \in C^{\#}$ such that

$$l(\bar{x}) \leqq l(x) \qquad (x \in S).$$

It is obvious that the set of properly minimal elements of S is always contained in the set of minimal elements of S.

Suppose there are two norms $p : E \to \mathbf{R}_+$ and $q : E \to \mathbf{R}_+$. We say that p is *equivalent* to q if there are two positive numbers m and M such that

$$mp(\cdot) \leqq q(\cdot) \leqq Mp(\cdot).$$

Thus p is equivalent to q if and only if p and q induce the same topology on E.

Let $(E, \|\cdot\|)$ be a real normed space, and let S^* denote the unit sphere of the dual space E^*. A subset K of E is called a Bishop-Phelps cone if there is some $l \in S^*$ and $\alpha \in (0, 1]$ such that

$$K = \{x \in E : l(x) \geqq \alpha \|x\|\}.$$

It is easy to see that K is a pointed closed convex cone. If K is not trivial, then K possesses the closed bounded base

$$B := \{x \in K : l(x) = 1\}.$$

But even in $\mathbf{R}^3$ with the Euclidean norm it is not true that every convex cone with a closed bounded base is a Bishop-Phelps cone. The reason for this phenomenon is that every base of a Bishop-Phelps cone in this space must be the convex hull of some ellipse. Thus the natural ordering cone $\mathbf{R}^3_+$ cannot be a Bishop-Phelps cone, since every base of it must be a triangle. On the other hand, $\mathbf{R}^3_+$ is a Bishop-Phelps cone if $\mathbf{R}^n$ is equipped with the l_1-norm. In the next section we will investigate the relation of Bishop-Phelps cones with respect to different norms.

3. Representation of cones as Bishop–Phelps cones.

DEFINITION 3.1. Let $(E, \|\cdot\|)$ be a normed space. Let C be a subset of E. We say that C is *representable* as a Bishop-Phelps cone, if there is some $l \in E^*$ and a norm $p : E \to \mathbf{R}_+$, which is equivalent to $\|\cdot\|$, such that

$$C = \{x \in E : p(x) \leqq l(x)\}.$$

In view of the examples of Bishop-Phelps cones at the end of § 2, we now are able to give a unified characterization of Bishop-Phelps cones generated by different norms.

THEOREM 3.2. *Let $(E, \|\cdot\|)$ be a real normed space. Let C be a nonempty subset of E with $C \neq \{0\}$. Then the following assertions are equivalent*:

(1) *C is representable as a Bishop-Phelps cone.*

(2) *C is a convex cone with a closed bounded base.*

Let us note here that every convex cone with a closed bounded base is closed and pointed [7, Prop. 3.8.3].

Proof. If C is representable as a Bishop-Phelps cone, then there is some $l \in E^*$ and a norm $p: E \to \mathbf{R}_+$, which is equivalent to $\|\cdot\|$ and such that

$$C = \{x \in E: p(x) \leqq l(x)\}.$$

It is easy to see that C is a convex cone. The continuity of l and p imply that C is closed. Let $B := \{x \in C: l(x) = 1\}$. Then B is a base (cf. [6, Lemma 3.3]). We have $p(x) \leqq 1$ for all $x \in B$. Since $\|\cdot\|$ and p are equivalent norms, the set B is bounded. This establishes (1) of Theorem 3.2.

Let C be a convex cone with a closed bounded base B_0. Since $0 \notin B_0$ there is some continuous linear functional, which strictly separates the convex sets $\{0\}$ and B_0. Thus we have some $l \in E^*$ and some $\kappa > 0$ such that

$$0 < \kappa \leqq l(b) \qquad (b \in B_0).$$

Let us consider the set $B := \{x \in C: l(x) = 1\}$. It is easy to verify that B is a closed base of C. Since B_0 is bounded there is some $M > 0$ such that $\|b\| \leqq M$ for all $b \in B_0$. We will show now that B is also bounded. Let x be an element of B. Then there are $\rho > 0$ and $b \in B_0$ such that $x = \rho b$. Thus we have

$$\rho = \rho/1 = \rho/l(x) = 1/l(b) \leqq \kappa^{-1}.$$

From this relation we derive the inequality

$$\|x\| = \rho\|b\| \leqq \kappa^{-1}M,$$

which shows that B is bounded. Now pick $\delta > 0$ such that $l(u) \leqq \frac{1}{2}$ for all $u \in U_\delta := \{x \in E: \|x\| \leqq \delta\}$. Let

$$F := \text{conv.}\,(-B \cup U_\delta \cup B).$$

Then F is convex balanced and absorbing. These properties of F imply that the corresponding *Minkowski* functional

$$p(x) := \inf\{t > 0: x \in tF\} \qquad (x \in E)$$

is a seminorm on E. Furthermore, F is the convex hull of a bounded set. Thus F is also bounded. Consequently, p is a norm on E. Since F is bounded, there is some $m > 0$ with $\|x\| \leqq m$ for all $x \in F$. From the convexity of the sets B, $-B$, and U_δ, it follows that each $x \in F$ has a representation

$$x = \lambda b + \mu u + \nu(-d)$$

for some elements $b, d \in B$, $u \in U_\delta$ and nonnegative numbers λ, μ, ν, that satisfy the condition $\lambda + \mu + \nu = 1$.

Now we will show that $p(\cdot)$ and $\|\cdot\|$ are equivalent. Let $x \in E$ with $\|x\| \leqq \delta$. Then we have $x \in U_\delta \subset F$, and this renders $p(x) \leqq 1$.

Now let $x \in E$ with $p(x) \leqq 1$. Then we get $x \in \bar{F}$. Let $M := \max\{m, \delta\}$. Thus we have $\|x\| \leqq M$. The two preceding conclusions can be summarized in the following inequality:

$$\delta p(x) \leqq \|x\| \leqq Mp(x) \qquad (x \in E).$$

This shows that $p(\cdot)$ and $\|\cdot\|$ are equivalent.

The proof will be complete if we can show that

$$C = \{x \in E : p(x) \leqq l(x)\}. \tag{1}$$

We will denote the right side of (1) by A. First let us show the inclusion $C \subset A$. Let $x \in C$. Then there is some $\lambda \geqq 0$ and some $b \in B$ such that $x = \lambda b$. Thus we have $x \in \lambda B \subset \lambda F$, and this renders $p(x) \leqq \lambda$. Consequently, we get

$$p(x) \leqq \lambda \cdot 1 = \lambda l(b) = l(\lambda b) = l(x)$$

and this implies $x \in A$. It remains to show the inclusion $A \subset C$. We need only to prove that any element $x \in E$ is contained in C if it satisfies the conditions $p(x) = 1$ and $l(x) \geqq 1$. So let $x \in E$ be an element with these properties. Since $p(x) = 1$, we have $x \in \tau F$ for all $\tau > 1$. Now let $(\tau_n)_{n \in \mathbf{N}}$ be a sequence of real numbers with $\tau_n > 1$ for all $n \in \mathbf{N}$ and $\lim_{n \to \infty} \tau_n = 1$. Since $x \in \tau_n F$ for each $n \in \mathbf{N}$ there is some element $b_n \in B$, $d_n \in B$, and $u_n \in U_\delta$ and nonnegative numbers λ_n, μ_n, ν_n with $\lambda_n + \mu_n + \nu_n = 1$ such that

$$x = \tau_n(\lambda_n b_n + \mu_n u_n + \nu_n(-d_n)). \tag{2}$$

We have $0 \leqq \lambda_n \leqq 1$, $0 \leqq \mu_n \leqq 1$, $0 \leqq \nu_n \leqq 1$. Then we can select convergent subsequences $(\lambda_{n_j})_{j \in \mathbf{N}}$, $(\mu_{n_j})_{j \in \mathbf{N}}$, $(\nu_{n_j})_{j \in \mathbf{N}}$. We denote the limits of these subsequences by λ, μ, and ν, respectively. Since $l(x) \geqq 1$ we derive from (2)

$$1 \leqq l(x) \leqq \tau_{n_j}(\lambda_{n_j} + \tfrac{1}{2}\mu_{n_j} - \nu_{n_j}). \tag{3}$$

If we let $j \to \infty$ in (3), we obtain the following inequality:

$$1 \leqq \lambda + \tfrac{1}{2}\mu - \nu. \tag{4}$$

Since $\lambda_{n_j} + \mu_{n_j} + \nu_{n_j} = 1$, as $j \to \infty$ we get

$$1 = \lambda + \mu + \nu. \tag{5}$$

Combining the inequality (4) with (5) leads to the inequality

$$\tfrac{1}{2}\mu + 2\nu \leqq 0.$$

Since μ and ν are nonnegative, we have $\mu = 0$ and $\nu = 0$. From this it follows that $\mu_{n_j} \to 0$, $\nu_{n_j} \to 0$, $\lambda_{n_j} \to 1$. Since U_δ and B are bounded, we have

$$\tau_{n_j}(\mu_{n_j} u_{n_j} + \nu_{n_j}(-d_{n_j})) \to 0 \qquad (j \to \infty).$$

So from (2) it follows that

$$\tau_{n_j} \lambda_{n_j} b_{n_j} \to x \qquad (j \to \infty).$$

Since $\tau_{n_j} \lambda_{n_j} \to 1$, it follows that $b_{n_j} \to x$. But b_{n_j} is contained in B. Consequently, we have $x \in \bar{B} = B \subset C$. Thus $x \in C$ and the desired conclusion has been shown. □

Now we will investigate convex cones in $\mathbf{R}^n$. In the sequel let $\mathbf{R}^n$ be equipped with the Euclidean norm.

LEMMA 3.3. *Every pointed closed convex cone in $\mathbf{R}^n$ admits a compact base.*

Proof. Since $\mathbf{R}^n$ is locally compact, a well-known theorem of Klee [7, Thm. 3.12.8] renders the assertion. □

The next assertion follows directly from Lemma 3.3 and Theorem 3.2.

THEOREM 3.4. *A convex cone in $\mathbf{R}^n$ is representable as a Bishop–Phelps cone if and only if it is closed and pointed.*

4. A generalization of the Arrow–Barankin–Blackwell Theorem. Now we are able to prove the previously announced results.

THEOREM 4.1. *Let $(E, \|\cdot\|)$ be a real normed space, partially ordered by a closed convex cone C with a closed bounded base. Let S be a nonempty convex subset of E, and let $\bar{x} \in S$ be a minimal element of S such that the set*

$$\hat{S} := \{x \in S: \|x - \bar{x}\| \leqq 1\}$$

is weakly compact. Then for every $\varepsilon > 0$ there is some $l_\varepsilon \in C^\#$ and some $x_\varepsilon \in S$ such that

$$l_\varepsilon(x_\varepsilon) = \min_{x \in S} l_\varepsilon(x) \quad \text{and} \quad \|x_\varepsilon - \bar{x}\| \leqq \varepsilon.$$

Proof. Let $\bar{x}$ be some minimal element of S with respect to C and choose $\varepsilon > 0$. Since C is a convex cone with a closed bounded base we can apply Theorem 3.2. Thus there is some $l \in E^*$ and a norm $p: E \to \mathbf{R}_+$ that is equivalent to $\|\cdot\|$ such that

$$C = \{x \in E: p(x) \leqq l(x)\}.$$

Since p is equivalent to $\|\cdot\|$ there are numbers $m > 0$, $M > 0$ with

$$m\|\cdot\| \leqq p(\cdot) \leqq M\|\cdot\|.$$

So $q := (1/m)p$ is also a norm equivalent to $\|\cdot\|$. Let

$$T := \{x \in S: q(x - \bar{x}) \leqq 1\}.$$

It is easy to check that T is a subset of $\hat{S}$. The set T is closed since q is equivalent to $\|\cdot\|$. Thus T is a closed convex subset of the weakly compact set $\hat{S}$, and therefore T is weakly compact also. Then T is also weakly compact in the space (E, q) since the weak topologies of (E, q) and $(E, \|\cdot\|)$ coincide. Another consequence of the equivalence of $\|\cdot\|$ and q is that $C^\#$ remains the same for both spaces $(E, \|\cdot\|)$ and (E, q). Let us define

$$\beta := \sup\left\{\frac{1}{m} l(x): x \in E, q(x) \leqq 1\right\}.$$

Since l is continuous in (E, q) and $C \neq \{0\}$ the inequality $1 \leqq \beta < \infty$ holds. Now we let $\alpha := (1/\beta) \in (0, 1]$ and $l_0 := (\alpha/m)l$. Then l_0 has norm one in the dual space of (E, q) and we have

$$C = \{x \in E: \alpha q(x) \leqq l_0(x)\}.$$

Thus C is a Bishop-Phelps cone in (E, q). Furthermore, T is a weakly compact set in (E, q).

We now may apply Jahn's result [5, Thm. 3.1] to the set S as a subset of the space (E, q): Then for $\varepsilon > 0$ there is some x_ε and some $l_\varepsilon \in C^\#$ such that

$$l_\varepsilon(x_\varepsilon) \leqq l(x) \qquad (x \in S)$$

and $q(\bar{x} - x_\varepsilon) \leqq \varepsilon$. Thus we have

$$\|x_\varepsilon - \bar{x}\| \leqq \frac{1}{m} p(x_\varepsilon - \bar{x}) = q(x_\varepsilon - \bar{x}) \leqq \varepsilon.$$

This completes the proof. □

In the remainder of this section we will present some special cases of Theorem 4.1.

In [3] Borwein has already proved that the set of minimal elements of some weakly compact convex set is contained in the weak closure of the set of properly minimal elements, if the ordering is induced by a cone with a compact base. So in a reflexive space the set of properly minimal elements is dense (with respect to the weak topology) in the set of minimal elements if the cone has a closed bounded base.

Jahn has shown in [5, Cor. 3.5] a result of similar type. He first considered a space ordered by some cone C with a closed bounded base. In this case he has proved that a minimal element of $\bar{x}$ of some weakly compact set can be approximated (with respect to the norm topology) by properly minimal elements, if $\bar{x}$ is also minimal with respect to some Bishop-Phelps cone containing C.

We can obtain both results as immediate consequences of the following corollary, which follows easily from Theorem 4.1.

COROLLARY 4.2. *Let $(E, \|\cdot\|)$ be a real normed space, partially ordered by a convex cone with a closed bounded base, and let S be a weakly compact convex subset of E. Then the set of properly minimal elements of S is dense (with respect to the norm topology) in the set of minimal elements.*

If $E = \mathbf{R}^n$, we derive from Theorem 4.1 and Lemma 3.3 the result of Bitran and Magnanti [2, Cor. 3.1].

COROLLARY 4.3. *Let $\mathbf{R}^n$ be partially ordered by a closed convex pointed cone. Let S be a closed convex subset of $\mathbf{R}^n$. Then the set of properly minimal elements of S is dense in the set of minimal elements of S.*

Since every locally convex Hausdorff space of finite dimensions is isomorphic to $\mathbf{R}^n$ (for a suitable integer n), Corollary 4.3 remains true for such spaces.

5. Example. In this section we will present an application of Corollary 4.2. We will consider an example from [6, Ex. 4.6].

5.1. Cooperative n player game. Let $X, E_1, \cdots, E_n$ be real locally convex Hausdorff spaces. Let each space E_i $(i = 1, \cdots, n)$ be ordered by a closed convex cone C_i $(i = 1, \cdots, n)$. Furthermore, let T be a nonempty subset of X. For each player there is given a mapping $P_i : T \to E_i$, which he tries to minimize on T.

As all players act exclusively cooperative this game can be viewed as a problem of vector optimization.

5.2. Problem of vector optimization. Let $E := E_1 \times E_2 \times \cdots \times E_n$ be the usual product space ordered by the convex cone $C := C_1 \times C_2 \times \cdots \times C_n$. Let $P : T \to E$ be the function $P(t) := (P_1(t), \cdots, P_n(t))$ $(t \in T)$.

In this setting we can reformulate the cooperative n player game: all points of $S := P(T)$, which are minimal with respect to C, must be localized. We now are interested in conditions that assure the Arrow-Barankin-Blackwell Theorem holds in E, if it holds in each E_i. Corollary 4.2 states that this theorem holds in each E_i, if C_i has a closed bounded base.

LEMMA 5.3. *If each of the convex cones $C_1, \cdots, C_n$ has some closed bounded base, then $C := C_1 \times C_2 \times \cdots \times C_n$ also has a closed bounded base.*

Proof. Let $B_1, \cdots, B_n$ be the closed bounded bases of $C_1, \cdots, C_n$, respectively. Consider the set

$$B := \bigcup_{\lambda_1, \cdots, \lambda_n \geq 0; \Sigma \lambda_i = 1} \lambda_1 B_1 \times \cdots \times \lambda_n B_n.$$

We easily check that B is a closed bounded base of C. □

So if $C_1, \cdots, C_n$ all have closed bounded bases, with the aid of Lemma 5.3 we can immediately apply Corollary 4.2 to the setting in § 5.2.

REFERENCES

[1] K. J. ARROW, E. W. BARANKIN, AND D. BLACKWELL, *Admissible points of convex sets*, in Contributions to the Theory of Games, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, 1953.

[2] G. R. BITRAN AND T. L. MAGNANTI, *The structure of admissible points with respect to cone dominance*, J. Optim. Theory Appl., 29 (1979), pp. 573–614.
[3] J. M. BORWEIN, *The Geometry of Pareto efficiency over cones*, Math. Operationsforsch. Statist. Ser. Optim., 11 (1980), pp. 235–248.
[4] R. B. HOLMES, *Geometric Functional Analysis and Its Applications*, Springer-Verlag, New York, Heidelberg, Berlin, 1975.
[5] J. JAHN, *A generalization of a theorem of Arrow–Barankin–Blackwell*, SIAM J. Control Optim., 26 (1988), pp. 999–1005.
[6] ———, *Mathematical Vector Optimization in Partially Ordered Linear Spaces*, Verlag Peter Lang, Frankfurt, 1986.
[7] G. JAMESON, *Ordered Linear Spaces*, Springer-Verlag, Berlin, Heidelberg, New York, 1970.
[8] W. RUDIN, *Functional Analysis*, Tata McGraw-Hill, New Delhi, THM edition, 1974, 7th reprint, 1982.
[9] W. SALZ, *Eine topologische Eigenschaft der effizienten Punkte konvexer Mengen*, Oper. Res. Verfahren, 23 (1976), pp. 197–202.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 402-422, March 1990

009

THE EXISTENCE OF CATCHING-UP OPTIMAL SOLUTIONS FOR A CLASS OF INFINITE HORIZON OPTIMAL CONTROL PROBLEMS WITH TIME DELAY*

DEAN A. CARLSON†

Abstract. The optimal control of a system whose states are governed by a nonlinear autonomous Volterra integrodifferential equation with unbounded time interval is considered. Specifically, it is assumed that the delay occurs only in the state variable. The results obtained extend those of Brock and Haurie [*Math. Oper. Res.*, 1 (1976), pp. 337-346] and Leizarowitz [*Math. Oper. Res.*, 10 (1985), pp. 450-461]. In particular, it is shown that (under appropriate hypotheses) catching-up optimal solutions asymptotically approach a unique optimal steady state, and thus enjoy the so-called "turnpike" property found in the economics literature. By combining this result with an associated optimal control problem, the desired existence result is obtained. Furthermore, it is remarked that, in addition to extending these earlier works to the time-delay case, the results presented below utilize convexity, seminormality, and growth hypotheses that in some cases are weaker than those encountered in the above-mentioned papers.

Key words. catching-up solutions, optimal control, integrodifferential equations

AMS(MOS) subject classifications. primary 49A10; secondary 90A

1. Introduction. The study of optimal control problems defined on infinite intervals has recently been a rapidly growing area of research. The primary area of application of these problems, although not the only one, concerns models of economic growth in which we search for a path of optimal capital accumulation. In these models it is shown that there exists an optimal asymptotic sustainable consumption in the economy. For a detailed introduction to optimal control problems of this type the reader is referred to the monograph of Carlson and Haurie [6].

It has long been recognized that time delays are important in formulating economic models. Indeed as early as 1935, Kalecki [14] introduced a class of such models described by differential-difference equations. These models were further investigated by Leontief [16] and others. More recently, models with infinite delay were formulated to describe optimal dynamic advertising. In particular, we refer to Hartl [12] for such a model as well as to the article of Hartl and Sethi [13].

In the present paper, we are concerned with the existence of catching-up optimal solutions for a class of models in which the states are governed by a nonlinear autonomous Volterra integrodifferential equation with infinite delay where the delay occurs only with respect to the state variable. The results we obtain extend the original results of Brock and Haurie [5], as well as the generalization given in Leizarowitz [15], for optimal control models whose states are governed by autonomous ordinary differential equations (see also Carlson and Haurie [6]).

With these remarks, the plan of the work presented below is as follows. In § 2, we introduce the model considered and indicate the basic hypotheses assumed throughout our work. Section 3 is devoted to summarizing several technical results concerning linear hereditary operators due to Marcus and Mizel [17] as well as presenting a general lower closure theorem that will be utilized in our presentation. The desired existence results are given in § 4 and we conclude our discussion in § 5

* Received by the editors March 10, 1988; accepted for publication (in revised form) May 24, 1989.

† Department of Mathematics, University of Toledo, Toledo, Ohio 43606. This research was supported by National Science Foundation grants DMS-8521465 and DMS-8700706.

by presenting several examples, including a version of the classical Ramsey model with a distributed time delay.

2. The basic model. We consider a system described by a Volterra integrodifferential equation of the form

$$\dot{x}(t) = f(x(t), u(t)) + \int_{-\infty}^{t} g(t-s)h(x(s))\, ds \quad \text{a.e. } t \geqq 0, \tag{2.1}$$

where $x: (-\infty, \infty) \to E^n$ is a bounded continuous function that is locally absolutely continuous on $[0, \infty)$ and statisfies the prescribed initial condition

$$x(s) = x_0(s) \quad \text{for all } s < 0, \tag{2.2}$$

where $x_0: (-\infty, 0] \to E^n$ is a given bounded continuous function, as well as the state constraints

$$x(t) \in X \quad \text{for } t \in (-\infty, \infty), \tag{2.3}$$

in which X is a closed subset of the n-dimensional Euclidean space E^n. The control function $u: [0, \infty) \to E^m$ is assumed to be Lebesgue measurable and satisfies the feedback control constraints

$$u(t) \in U(x(t)) \quad \text{a.e. } 0 \leqq t, \tag{2.4}$$

where $U: X \to 2^{E^m}$ is a point to set mapping with closed graph $M = \{(x, u): x \in X \text{ and } u \in U(x)\}$.

As regards the functions f, g, and h, we assume that $f: M \to E^n$ and $h: X \to E^p$ are both continuous and that $g = (g_{ij})_{n \times p}$ is an $n \times p$ matrix function defined for $t \geqq 0$ with entries satisfying

$$\begin{aligned} &\text{(i)} \quad \int_0^{\infty} |g_{ij}(t)|\, dt < \infty, \\ &\text{(ii)} \quad \int_0^{\infty} t|g_{ij}(t)|\, dt < \infty, \\ &\text{(iii)} \quad \sum_{m=1}^{\infty} \|g_{ij}\|_m^{\infty} < \infty, \end{aligned} \tag{2.5}$$

where $\|g_{ij}\|_m^{\infty}$ is the essential supremum of g_{ij} restricted to the interval $[m-1, m]$, $m = 1, 2, \cdots$. We remark that the assumption (2.5)(iii) given above implies (2.5)(i), but we have included both for definiteness.

The performance of the above control system is described for any positive time T by the integral functional

$$J_T(x, u) = \int_0^T f^0(x(s), u(s))\, ds, \tag{2.6}$$

where $f^0: M \to E^1$ is a given lower semicontinuous function.

With this notation, we give the following definition.

DEFINITION 2.1. A bounded continuous function $x: (-\infty, \infty) \to E^n$ will be called a *trajectory* if x is locally absolutely continuous on $[0, \infty)$ and if there exists a Lebesgue measurable function (referred to as a control) $u: [0, \infty) \to E^m$ such that the pair $\{x, u\}$ satisfies (2.1), (2.3), (2.4), and the map $t \to f^0(x(t), u(t))$ is locally Lebesgue integrable on $[0, \infty)$. If in addition, the trajectory x satisfies the prescribed initial condition (2.2), we will call x an admissible trajectory and u an admissible control.

For brevity we let A denote the set of all trajectory-control pairs $\{x, u\}$ and let $A_0 \subset A$ be the set of all admissible pairs.

Since the system described above is defined for all $t \geqq 0$, we are primarily concerned with the performance over the entire interval $[0, \infty)$. We do not, however, assume a priori that the performance criterion $J_T(x, u)$ has a finite limit as $T \to \infty$. This necessitates the need to consider a weaker notion of optimality. For our treatment we restrict our attention to the following definitions.

DEFINITION 2.2. An admissible pair $\{x^*, u^*\} \in A_0$ is called

(i) *Strongly optimal* if $J_\infty(x^*, u^*) = \lim_{T\to\infty} J_T(x^*, u^*)$ is finite and if for every $\{x, u\} \in A_0$ we have

$$J_\infty(x^*, u^*) \leqq \lim_{T\to\infty} \inf J_T(x, u); \tag{2.7}$$

(ii) *Catching-up* (or overtaking) *optimal* if for each $\varepsilon > 0$ and pair $\{x, u\} \in A_0$, there exists $\tau = \tau(\varepsilon, x, u) \geqq 0$ such that for all $T \geqq \tau$ we have

$$J_T(x^*, u^*) \leqq J_T(x, u) + \varepsilon. \tag{2.8}$$

Remark 2.1. The notion of strong optimality given above is, of course, the traditional concept of a minimizer. On the other hand, catching-up optimality is a weaker concept that was introduced by von Weiszäcker [20] in 1965. The term overtaking optimality was apparently first used in Brock and Haurie [5]. For a detailed treatment of this concept of optimality as well as several others we refer the reader to [6].

As is usual in discussions of existence of optimal solutions for optimal control problems, it is necessary to place certain convexity and growth hypotheses on the model. These conditions are needed to ensure that appropriate lower semicontinuity and compactness properties hold. The assumptions we require are described as follows.

(A1) For each $x \in X$, the set $\tilde{\mathbb{Q}}(x)$ given by

$$\tilde{\mathbb{Q}}(x) = \{(z^0, z) \in E^{1+n}: z^0 \geqq f^0(x, u), z = f(x, u), u \in U(x)\} \tag{2.9}$$

is a nonempty closed, convex set that satisfies the upper semicontinuity condition property (K) given as

$$\tilde{\mathbb{Q}}(x) = \bigcap_{\delta>0} \operatorname{cl}[\cup \{\tilde{\mathbb{Q}}(y): |y - x| < \delta\}] \tag{2.10}$$

where $|\cdot|$ denotes the usual Euclidean norm on E^n.

(A2) We assume that for each $\varepsilon > 0$ there exists $c_\varepsilon > 0$ such that for all $(x, u) \in M$ we have

$$|f(x, u)| + |h(x)| \leqq c_\varepsilon + \varepsilon f^0(x, u). \tag{2.11}$$

Remark 2.2. The conditions placed on the sets $\tilde{\mathbb{Q}}(x)$ described above are standard and it is well known that for each $T > 0$ they guarantee, in the ordinary case (i.e., $h(x) = 0$), that the functional $(x, u) \to J_T(x, u)$ is lower semicontinuous on A_0 with respect to the weak topology in $\mathrm{AC}([0, T]; E^n)$, the space of absolutely continuous functions, placed on the set of admissible trajectories. (That is, the topology of pointwise convergence of initial conditions and weak L_1-convergence in the derivatives of x.) The growth condition given in (A2) provides for the equi-absolute integrability of the derivatives of the admissible trajectories on $[0, T]$ in a minimizing sequence, and consequently (by the Dunford–Pettis criterion), gives the requisite compactness conditions. We further remark that in the ordinary differential equation case this growth

condition is weaker than the growth condition used in Leizarowitz [15], but it is equivalent to the classical growth hypothesis of Nagumo and Tonelli, which is referred to in Proposition 3.2 of [15]. For a complete discussion of these matters see Cesari [9, § 10.4].

To conclude our description of the model, we introduce the optimal steady-state problem. As is usual in the treatment of autonomous infinite horizon optimal control problems, it will be established that the catching-up optimal solution, which we exhibit, will converge to the optimal steady state. In the economics literature this is commonly called a turnpike property. Such a property was first introduced by Samuelson [18] (see also Cass [8]) in the context of optimal economic growth. We now describe this steady-state problem.

(A3) We assume that the optimal steady-state problem (OSSP) described as

(i) minimize $f^0(x, u)$

subject to

(2.12) (ii) $0 = f(x, u) + \left(\int_0^\infty g(s)\, ds \right) h(x),$

(iii) $x \in X,$

(iv) $u \in U(x)$

has a solution $(\bar{x}, \bar{u}) \in E^{n+m}$, with $\bar{x}$ uniquely determined.

We further assume that there exists $\bar{p} \in E^n$ such that the lower semicontinuous function $L: M \to E^1$ given by

$$L(x, u) = f^0(x, u) - f^0(\bar{x}, \bar{u}) + \left\langle \bar{p}, f(x, u) + \left(\int_0^\infty g(s)\, ds \right) h(x) \right\rangle \tag{2.13}$$

is nonnegative, where $\langle \cdot, \cdot \rangle$ denotes the usual inner product in E^n.

Remark 2.3. The OSSP described above plays an analogous role for the nondelay infinite horizon optimal control problem described by

$$\text{minimize} \int_0^{+\infty} f^0(x(t), u(t))\, dt$$

subject to

$$\dot{x}(t) = f(x(t), u(t)) + \left(\int_0^\infty g(s)\, ds \right) h(x(t)) \quad \text{a.e. } t \geqq 0,$$

$$x(0) = x_0(0),$$

$$x(t) \in X \quad \text{from } t \in (-\infty, \infty),$$

and

$$u(t) \in U(x(t)) \quad \text{a.e. } [0, \infty).$$

In fact, under essentially the same hypotheses as those utilized here, we can establish the existence of a catching-up optimal solution, say $\{\hat{x}, \hat{u}\}$, for the nondelay case (see, e.g., Brock and Haurie [5] and Leizarowitz [15]) by appealing to established results. We refer the reader to these results that are outlined in Carlson and Haurie [6]. In particular, we remark that in all these results it is established that

$$\lim_{t \to \infty} \hat{x}(t) = \bar{x}.$$

As we will see in § 4 this asymptotic convergence property also holds for the delay case considered here. In this way we see that for large $t>0$, the optimal trajectories for both systems are close together.

Concerning the function L we have the following elementary result.

PROPOSITION 2.1. *Under the hypotheses placed on X, U, f^0, f, h, and g, if the sets $\tilde{\mathbb{Q}}(x)$, given by* (2.9), *satisfy the conditions outlined in* (A1) *we have that the sets $\tilde{\mathbb{Q}}_L(x)$ defined for $x \in X$ by*

$$\tilde{\mathbb{Q}}_L(x) = \{(z^0, z): z^0 \geqq L(x, u),\ z = f(x, u),\ u \in U(x)\}$$

enjoy the same properties. Furthermore, if f, h, and f^0 satisfy the growth condition (A2), *the same holds for f, h, and L. That is, for each $\varepsilon > 0$ there exists $c_\varepsilon > 0$ so that*

$$|f(x, u)| + |h(x)| \leqq c_\varepsilon + \varepsilon L(x, u)$$

for all $(x, u) \in M$.

Proof. The convexity and upper semicontinuity properties of the sets $\tilde{\mathbb{Q}}_L(x)$ are an easy consequence of the fact that for each $x \in X$ the affine mapping $\Gamma_x : \tilde{\mathbb{Q}}(x) \to \tilde{\mathbb{Q}}_L(x)$ defined by

$$\Gamma_x(z^0, z) = \left(z^0 - f^0(\bar{x}, \bar{u}) + \left\langle \bar{p}, z + \left(\int_0^\infty g(s)\, ds \right) h(x) \right\rangle, z \right)$$

is one-to-one and onto. We leave the details of this argument to the reader.

To establish the growth condition we let $\varepsilon > 0$ be given, let $K = \max [|\bar{p}|, |\bar{p}| \int_0^\infty g(s)\, ds|]$, and observe that for each $(x, u) \in M$ we have

$$\begin{aligned} f^0(x, u) &= L(x, u) + f^0(\bar{x}, \bar{u}) - \left\langle \bar{p}, f(x, u) + \left(\int_0^\infty g(s)\, ds \right) h(x) \right\rangle \\ &\leqq L(x, u) + f^0(\bar{x}, \bar{u}) + K[|f(x, u)| + |h(x)|]. \end{aligned}$$

Let $\eta > 0$ be chosen so that $\eta < \min [\varepsilon/(1+\varepsilon K), 1/K]$ if $K \neq 0$ and $\eta = \varepsilon$ if $K = 0$. From (A2) there exists $c_\eta > 0$ such that (2.11) holds with $\varepsilon = \eta$. This implies that for all $(x, u) \in M$,

$$|f(x, u)| + |h(x)| \leqq c_\eta + \eta[L(x, u) + f^0(\bar{x}, \bar{u}) + K[|f(x, u)| + |h(x)|]]$$

or, equivalently,

$$(1 - \eta K)[|f(x, u)| + |h(x)|] \leqq [c_\eta + \eta f^0(\bar{x}, \bar{u})] + \eta L(x, u).$$

From our choice of η it follows that $(1 - \eta K)^{-1}\eta < \varepsilon$ and that $1 - \eta K > 0$. Therefore, we obtain

$$\begin{aligned} |f(x, u)| + |h(x)| &\leqq (1 - \eta K)^{-1}(c_\eta + \eta f^0(\bar{x}, \bar{u})) + \frac{\eta}{1 - \eta K} L(x, u) \\ &\leqq (1 - \eta K)^{-1}(c_\eta + \eta f^0(\bar{x}, \bar{u})) + \varepsilon L(x, u). \end{aligned}$$

The desired conclusion follows by choosing $c_\varepsilon > 0$ satisfying

$$c_\varepsilon > (1 - \eta K)^{-1}(c_\eta + \eta f^0(\bar{x}, \bar{u})).$$

3. Linear hereditary operators and a lower closure theorem. To establish our results we will need certain properties of the linear integral operator G defined by

$$(3.1) \qquad (Gy)(t) = \int_{-\infty}^{t} g(t-s)y(s)\, ds = \int_0^\infty g(s)y(t-s)\, ds, \qquad -\infty < t < \infty,$$

in which g is an $n \times p$ matrix function satisfying the condition (2.5)(iii). Clearly, the operator G is not well defined for all choices of functions y. However, operators of

this type have been thoroughly examined in Marcus and Mizel [17], who chose as a domain for this operator the space $N_1((-\infty, \infty); E^p)$ consisting of all locally integrable functions $y \in L^1_{\text{loc}}((-\infty, \infty); E^p)$ with the following property:

$$|||y||| = \sup_{-\infty < s < \infty} \left\{ \int_{s-1}^{s} |y(t)|\, dt \right\} < \infty. \tag{3.2}$$

It is easy to see that $|||\cdot|||$ defines a norm on the space $N_1((-\infty, \infty); E^p)$. With this notation we summarize the results from Marcus and Mizel [17] that we require in the following theorem.

THEOREM 3.1. *Let g be an $n \times p$ matrix function on $[0, \infty)$ satisfying* (2.5)(iii) *and define the operator G on $N_1((-\infty, \infty); E^p)$ by* (3.1). *The following properties hold.*

(A) *For every $y \in N_1((-\infty, \infty); E^p)$ the integral $(Gy)(t)$, defined by* (3.1) *for each $t \in (-\infty, \infty)$, exists as a Lebesgue integral.*

(B) *For every $y \in N_1((-\infty, \infty); E^p)$, the function $G(y): (-\infty, \infty) \to E^n$ is continuous.*

(C) *If $\{y_k\}_{k=1}^{\infty} \subset N_1((-\infty, \infty), E^p)$ is a norm bounded sequence (i.e., there exists $R > 0$ so that $|||y_k||| \leqq R$ for all k) such that for some $y \in N_1((-\infty, \infty); E^p)$ the sequence $\{y_k|_{(\alpha,\beta)}\}_{k=1}^{\infty}$ converges strongly in $L^1(\alpha, \beta)$ to y for all real numbers α, β, $\alpha < \beta$; then the sequence $\{G(y_k)\}_{k=1}^{\infty}$ converges uniformly on compact subsets of $(-\infty, \infty)$ to $G(y)$.*

(D) *If $D \subset N_1((-\infty, \infty): E^p)$ is a norm bounded subset such that D is uniformly integrable on $(-T, T)$ for every $T > 0$, then the set $G(D) = \{G(y): y \in D\}$ is precompact in $C([-T, T]: E^n)$, the space of continuous functions on $[-T, T]$, for all $T > 0$.*

Proof. The proof of the above results are found in Lemmas 7.5 and 7.10 of [17].

Remark 3.1. In [17] the kernel g is assumed to satisfy stronger hypotheses than those indicated above (see [17, p. 21]). However, a careful examination of the proofs for the results given above shows that this additional hypothesis is not required. On the other hand, this additional hypothesis is needed to prove all the results given in [17, Lemma 7.5]. We also note that the proof of part (D) above given in [17, Lemma 7.10] is only concerned with $D \subset N((-\infty, 0); E^p)$ but a straightforward modification of their argument gives the result for the case considered here.

In addition to the above theorem, we will need the following lower closure theorem. For the applications of this result in this paper, the sets G_k below are subsets of the time axis $(-\infty, \infty)$.

THEOREM 3.2. *Let G be a σ-finite measure space, $G = \bigcup_{k=1}^{\infty} G_k$ where $G_k \subset G_{k+1}$ and* meas $(G_k) < \infty$, *let $X \subset E^n$ be closed, and let $R: X \to 2^{E^{1+r}}$ be a given set-valued map that is closed and convex valued and, in addition, satisfies the upper semicontinuity property (K) given by* (2.10). *Furthermore, assume that there exists measurable functions $\eta_k: G_k \to E^1$, $\lambda_k: G_k \to E^1$, $\lambda: G \to E^1$. $\xi_k: G_k \to E^n$, $\xi: G \to E^n$, $x_k: G_k \to E^n$, and $x: G \to E^n$, $k = 1, 2, \cdots$ satisfying the following conditions*:

(i) *For each index k, and almost all $t \in G_k$, $x_k(t) \in X$ and $(\eta_k(t), \xi_k(t)) \in R(x_k(t))$.*

(ii) *$x_k(t) \to x(t)$ pointwise, $\xi_k \to \xi$ weakly in $L^1_{\text{loc}}(G; E^n)$, and $\lambda_k \to \lambda$ weakly in $L^1_{\text{loc}}(G; E^1)$ as $k \to \infty$.*

(iii) *$\lambda \in L^1(G; E^1)$, $\eta_k(t) \geqq \lambda_k(t)$ almost everywhere in G_k, and $-\infty < i = \liminf_{k \to \infty} \int_{G_k} \eta_k(t)\, dt < +\infty$.*

Then there exists $\eta \in L^1(G; E^1)$ such that

$$x(t) \in X \quad \text{and} \quad (\eta(t), \xi(t)) \in R(x(t)) \quad \text{a.e. } t \in G$$

and

$$-\infty < \int_G \eta(t)\, dt \leqq i.$$

Remark 3.2. The above result, in a control-theoretic format, is originally due to Baum [4], in which the upper semicontinuity property (K) is replaced by property (Q). In terms of orientor fields (i.e., the above format) Baum's result was given by Bates [3]. Finally, the assumption property (Q) was replaced by property (K) in Balder [2].

4. Catching-up optimal solutions. With the notation and hypotheses given above we now address the problem of the existence of catching-up optimal solutions. We begin by investigating the asymptotic convergence properties of certain admissible trajectories to the optimal steady state $\bar{x}$. Following Leizarowitz [15], we let $\mathcal{F}$ denote the set of all trajectories $x: E^1 \to E^n$ satisfying

$$L(x(t), u(t)) = 0 \quad \text{a.e. } t > 0 \tag{4.1}$$

where $u: [0, \infty) \to E^n$ is a measurable control function corresponding to the trajectory x (see Definition 2.1). We observe that the optimal steady state $x(t) = \bar{x}$ is a trajectory such that $\mathcal{F}$ is nonempty. As regards to $\mathcal{F}$, we make the following assumption.

(A4) For each $\varepsilon > 0$ there exists $t_\varepsilon > 0$ such that for all $t \geqq t_\varepsilon$ and all $x \in \mathcal{F}$, $|x(t) - \bar{x}| < \varepsilon$.

Concerning Assumption (A4) we remark that it corresponds to property (S) in Leizarowitz [15] and to property $(\mathcal{C})$ of Carlson, Haurie, and Jabrane [7]. As stated, this assumption is difficult to verify and will not hold in general. In the ordinary differential equation case (i.e., $h(x) \equiv 0$) it has been shown in [15] that under suitable convexity conditions this condition is generic. The infinite-dimensional nature of the integrodifferential equation model considered here precludes such a possibility in our situation. Consequently, we content ourselves by presenting explicit conditions that imply this assumption. To this end we introduce the following alternative assumption.

(A4′) For every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ so that if $x \in X$ satisfies $|x - \bar{x}| > \varepsilon$, then $L(x, u) \geqq \delta$ for all $u \in U(x)$.

LEMMA 4.1. *If* (A4′) *holds, then* (A4) *holds.*

Proof. We assume that (A4′) holds and let $x: E^1 \to E^n$ be a trajectory (see Definition 2.1) so that $L(x(t), u(t)) = 0$ almost everywhere on $[0, \infty)$ (i.e., $x \in \mathcal{F}$). We now show that $x(t) \equiv \bar{x}$ on $[0, \infty)$. Indeed, if there exists $\tau \geqq 0$ for which $x(\tau) \neq \bar{x}$, the continuity of x allows us to assert the existence of an $\varepsilon_0 > 0$ and $\alpha > 0$ so that for all $t \in [\tau, \tau + \alpha)$ we have $|x(t) - \bar{x}| > \varepsilon_0$. However, by (A4′) there exists $\delta_0 > 0$ so that $L(x(t), u) > \delta_0$ for all $u \in U(x(t))$ and almost all $t \in [\tau, \tau + \alpha)$. Clearly, this is a contradiction. Therefore $x(t) \equiv \bar{x}$ on $[0, \infty)$ and thus, since $x \in \mathcal{F}$ was arbitrary, Assumption (A4) holds since for each $\varepsilon > 0$ we can choose $t_\varepsilon = 0$.

The assumption (A4′) appears in earlier works concerning infinite horizon optimal control (see, e.g., Carlson and Haurie [6, Chap. 4] and Brock and Haurie [5]), where its role is analogous to (A4). In fact, it was not until Leizarowitz [15] that (A4′) was replaced by (A4).

A weaker condition than (A4′), which is still sufficient for (A4) to hold, is

(A4″) The optimal steady state $\bar{x}$ is uniquely determined and if $L(x, u) = 0$, then $x = \bar{x}$.

It is easy to see that (A4′) implies (A4″). Moreover, it is also easy to see that Lemma 4.1 holds with (A4″) replacing (A4′).

The implications of (A4′) (or (A4″)) differ slightly from the ordinary differential equation case. In particular, (A4′) (or (A4″)) implies $\mathcal{F} = \{\bar{x}\}$ in the nondelay case. In the delay case considered here it is possible (under (A4′) or (A4″)) that there exists $x \in \mathcal{F}$ that is not identically equal to $\bar{x}$ on $(-\infty, \infty)$. This is demonstrated in the following example. However, we remark that we still essentially have $\mathcal{F}$ as the singleton $\mathcal{F} = \{\bar{x}\}$, since given $x \in \mathcal{F}$ with $x(t) \neq \bar{x}$ for $t < 0$ we must have (under (A4′) or (A4″)) that $x(t) = \bar{x}$ on $[\infty, \infty)$.

Example 4.1. We consider the control system

$$\dot{x}(t) = u(t) + \int_{-\infty}^{t} 2e^{-(t-s)}x(s)\,ds \quad \text{a.e. on } [0, \infty),$$

$$x(s) = \varphi(s) \quad \text{for all } s \leqq 0,$$

$$x(t) \in [-1, 1] \quad \text{on } [0, \infty),$$

$$u(t) \in [-1, 1] \quad \text{a.e. on } [0, \infty).$$

Here we assume that $\varphi : (-\infty, 0] \to [-1, 1]$ is a fixed given initial function. For the objective functional we take

$$J_T(x, u) = \int_0^T [2x(t)(4x(t)^2 + 6x(t) - 9) - 9u(t)]\,dt$$

for $T \geqq 0$. The corresponding optimal steady-state problem becomes

$$\text{minimize } \{2x(4x^2 + 6x - 9) - 9u : 0 = u + 2x,\ x \in [-1, 1],\ u \in [-1, 1]\}.$$

It is a straightforward argument to show that (A3) holds with $\bar{x} = 0$, $\bar{u} = 0$, and $\bar{p} = 9$. Thus,

$$L(x, u) = 2x(4x^2 + 6x - 9) - 9u + 9(2x + u) = 2x(4x^2 + 6x) = 4x^2(2x + 3)$$

and since $(x, u) \in [-1, 1] \times [-1, 1]$ we have

$$L(x, u) \geqq 4x^2,$$

so that if $|x - \bar{x}| = |x| > \varepsilon$ we have, upon choosing $\delta = \varepsilon^2/4$,

$$L(x, u) \geqq \delta.$$

Consequently, the assumption (A4′) holds. Clearly, $x(t) \equiv \bar{x} = 0$ is in $\mathcal{F}$. We now show that $\hat{x} : (-\infty, \infty) \to [-1, 1]$, defined by

$$\hat{x}(t) = \begin{cases} (\frac{1}{2})(e^t - 1) & \text{for } t < 0, \\ 0 & \text{for } t \geqq 0, \end{cases}$$

is also in $\mathcal{F}$. Indeed, we observe that by taking $\hat{u} : [0, \infty) \to [-1, 1]$ to be

$$\hat{u}(t) = \tfrac{1}{2}e^{-t},$$

we have for all $t \geqq 0$ that

$$\hat{u}(t) + \int_{-\infty}^{t} 2e^{-(t-s)}\hat{x}(s)\,ds = \frac{1}{2}e^{-t} + \int_{-\infty}^{0} 2e^{-(t-s)}\left(\frac{1}{2}(e^s - 1)\right) ds$$

$$= \frac{1}{2}e^{-t} - \frac{1}{2}e^{-t} = 0.$$

This implies that $\hat{x}$ is indeed a trajectory with $L(\hat{x}(t), \hat{u}(t)) = 0$ for all $t \geqq 0$ as desired.

Condition (A4′) is finite-dimensional in nature and as a consequence is easier to verify than (A4). To present general conditions that, on the model we consider, are sufficient for (A4′) to hold requires stronger convexity hypotheses. The following proposition gives conditions that permit us to conclude that (A4′) holds.

PROPOSITION 4.1. *Suppose that* $M=\{(x, u): x\in X, u\in U(x)\}$ *is compact and convex with nonempty interior, that* $f^0: M\to E^1$ *is strictly convex, lower semicontinuous and bounded below, and that* $F: M\to E^n$, *defined by*

$$F(x, u)=f(x, u)+\left(\int_0^\infty g(s)\,ds\right)h(x),$$

is continuous and concave (i.e., each component F_i, $F=(F_1, \cdots, F_n)$ *is a concave function). Furthermore, suppose the following*:

(i) *There exists* $(x, u)\in M$ *so that* $F_i(x, u)>0$ *for* $i=1, 2, \cdots, n$; *and*

(ii) *If* $(x, u)\in M$ *is such that* $F_i(x, u)\geqq 0$, $i=1, 2, \cdots, n$, *then there exists* $v\in U(x)$ *so that* $F(x, v)=0$ *and* $f^0(x, v)\leqq f^0(x, u)$.

Then (A3) *and* (A4′) *hold.*

Proof. By the classical Weierstrass Theorem it is clear that the optimization problem

$$\underset{(x,u)\in M}{\text{minimize}}\{f^0(x, u): 0\leqq F_i(x, u), 1\leqq i\leqq n\}$$

has a solution $(\bar{x}, \bar{u})$. Moreover, since f^0 is strictly convex, $(\bar{x}, \bar{u})$ is unique. Condition (i) is Slater's constraint qualification for the above convex program. Consequently, there exists a vector $\bar{p}=(\bar{p}_1, \cdots, \bar{p}_n)\in E^n$, $\bar{p}_i\geqq 0$, so that

$$f^0(x, u)+\langle \bar{p}, F(x, u)\rangle \geqq f^0(\bar{x}, \bar{u}),$$

for all $(x, u)\in M$. Moreover, from (ii) it follows that $F(\bar{x}, \bar{u})=0$ so that we clearly have that (A3) holds. To check (A4′) we proceed by contradiction. That is, we assume there exists $\varepsilon_0>0$ and a sequence of points $\{(x_n, u_n)\}_{n=1}^\infty$ in M so that $|x_n-\bar{x}|>\varepsilon_0$ but

$$\begin{aligned}0&\leqq L(x_n, u_n)\\ &=f^0(x_n, u_n)-f^0(\bar{x}, \bar{u})+\left\langle p, f(x_nu_n)+\left(\int_0^\infty g(s)\,ds\right)h(x_n)\right\rangle<\frac{1}{n}.\end{aligned}$$

The compactness of M allows us to choose a subsequence, say still $\{(x_n, u_n)\}_{n=1}^\infty$, which converges to some point $(\hat{x}, \hat{u})\in M$. This implies

$$0\geqq \liminf_{n\to\infty} L(x_n, u_n)\geqq L(\hat{x}, \hat{u})\geqq 0,$$

giving us $L(\hat{x}, \hat{u})=0$. However, the map $(x, u)\to f^0(x, u)+\langle \bar{p}, F(x, u)\rangle$, $(x, u)\in M$, is strictly convex so that $(\bar{x}, \bar{u})$ is the unique minimizer. Thus we must have $(\hat{x}, \hat{u})=(\bar{x}, \bar{u})$, which is clearly a contradiction since $|\hat{x}-\bar{x}|=\lim_{n\to\infty}|x_n-\bar{x}|>\varepsilon_0$. Thus, (A4′) holds.

Remark 4.1. The above proposition is not new and analogues of this result, for the ordinary differential equation case, appear in earlier works (see, e.g., Carlson and Haurie [6, Lemma 4.4]). While this result requires f^0 to be strictly convex on M we remark that this is only a sufficient condition for (A4′) and is not necessary. Indeed, this convexity condition is not satisfied in Example 4.1. To see this consider the points $(-1, 0)$ and $(-\frac{1}{2}, 0)$ and observe for $\lambda=\frac{1}{2}$ that we have

$$f^0\left(\lambda(-1)+(1-\lambda)\left(-\frac{1}{2}\right), 0\right)=\frac{135}{8}\geqq\frac{1}{2}f^0(-1, 0)+\frac{1}{2}f^0\left(-\frac{1}{2}, 0\right)=\frac{33}{2},$$

which violates the definition of convexity.

With this discussion concerning hypothesis (A4) in hand, we continue our presentation with the following result.

PROPOSITION 4.2. *Let $X \subset E^n$ be closed, let $U: X \to 2^{E^m}$ be a set-valued map with closed graph M, let $f^0: M \to E^1$ be lower semicontinuous, let $f: M \to E^n$ and $h: X \to E^p$ be continuous, and let g be an $n \times p$ matrix function satisfying* (2.5). *Furthermore, assume that* (A1)-(A4) *hold. If $\{x, u\} \in A_0$ is such that*

$$\int_0^\infty L(x(t), u(t))\, dt < \infty, \tag{4.2}$$

then $\lim_{t\to\infty} x(t) = \bar{x}$.

Proof. We proceed by contradiction and suppose that $\{x, u\}$ is as above, but $\lim_{t\to\infty} x(t) \neq \bar{x}$. This implies there exists an $\varepsilon_0 > 0$ and times t_k, $k = 1, 2, \cdots$, increasing to positive infinity such that $|x(t_k) - \bar{x}| > \varepsilon_0$ for all k. As a consequence of (A4) there exists $t_0 > 0$ such that for all $t \geqq t_0$ we have $|s(t) - \bar{x}| < \varepsilon_0/2$ for all trajectories $s \in \mathcal{F}$. Define the sequences of functions $x_k: (-\infty, \infty) \to E^n$ and $u_k: (t_0 - t_k, \infty) \to E^m$ by the formulas

$$x_k(t) = x(t + t_k - t_0), \qquad u_k(t) = u(t + t_k - t_0).$$

Clearly, we have for all $k = 1, 2, \cdots$, and almost all $t \geqq t_0 - t_k$ that

$$\dot{x}_k(t) = f(x_k(t), u_k(t)) + \int_{-\infty}^{t} g(t-s)h(x_k(t))\, ds,$$

$$u_k(t) \in U(x_k(t)), \qquad x_k(t) \in X.$$

We further note that for each $k = 1, 2, \cdots$, x_k satisfies the "initial condition"

$$x_k(s) = x_0(s + t_k - t_0) \quad \text{for all } s \leqq t_0 - t_k.$$

As we will see, since $(t_0 - t_k, \infty)$ tends to $(-\infty, \infty)$ as $k \to \infty$, this fact is not required for our proof. Moreover, since by definition $t \to x(t)$ is bounded, the sequence $\{x_k\}_{k=1}^\infty$ is uniformly bounded and satisfies $x_k(t_0) = x(t_k)$. From the above it is evident that x_k is a trajectory corresponding to the control function u_k (see Definition 2.1). In addition, we observe that for any $T > 0$, and all k sufficiently large

$$\lim_{k\to\infty} \int_{-T}^{T} L(x_k(t), u_k(t)) = \lim_{k\to\infty} \int_{t_k - t_0 - T}^{t_k - t_0 + T} L(x(t), u(t))\, dt = 0,$$

and thus by the growth condition (A2) applied to f, h, and L (see Proposition 2.1) we have that the sequences of functions $z_k(t) = f(x_k(t), u_k(t))$ and $y_k(t) = h(x_k(t))$, $k = 1, 2, \cdots$, are equi-absolutely integrable on $[-T, T]$ for all $T > 0$. Thus, by a standard diagonalization process there exists locally integrable functions $z: (-\infty, \infty) \to E^n$ and $y: (-\infty, \infty) \to E^p$ and subsequences, say still $\{z_k\}$ and $\{y_k\}$, that converge weakly to z and y, respectively, in $L^1_{\text{loc}}((-\infty, \infty), E^n)$ and $L^1_{\text{loc}}((-\infty, \infty), E^p)$. In addition, as the sequence $\{x_k\}$ is bounded and $h: X \to E^p$ is continuous, there exists $K > 0$ so that for all $t \in (-\infty, \infty)$ and all $k = 1, 2, \cdots$

$$|y_k(t)| = |h(x_k(t))| \leqq K.$$

Therefore, as a consequence of Theorem 3.1(D), we can further assume (by extracting another subsequence by diagonalization) that there exists a continuous function

$r:(-\infty,\infty)\to E^n$ such that the sequence $\{G(y_k)\}_{k=1}^{\infty}$, defined for $k=1, 2, \cdots$ by

$$G(y_k)(t)=\int_{-\infty}^{t} g(t-s)y_k(s)\,ds, \qquad t\in(-\infty,\infty),$$

converges uniformly on compact subsets of $(-\infty,\infty)$ to r. That is,

$$r(t)=\lim_{k\to\infty} G(y_k)(t)=\lim_{k\to\infty}\int_{-\infty}^{t}(g(t-s)y_k(s))\,ds. \tag{4.3}$$

Combining the above sequences we observe that for all $k=1, 2, \cdots$

$$\dot{x}_k(t)=z_k(t)+G(y_k)(t) \quad \text{a.e. } t_0-t_k\leqq t;$$

and we conclude that the sequence $\{\dot{x}_k\}_{k=1}^{\infty}$ converges weakly in $L^1_{\text{loc}}((-\infty,\infty);E^n)$ to the locally integrable function $t\mapsto z(t)+r(t)$. Also, since $\{x_k(t_0)\}_{k=1}^{\infty}$ is bounded, we can assume our subsequence has been chosen so that $\lim_{k\to\infty} x_k(t_0)=\hat{x}$, where $\hat{x}\in X$. Define $\hat{x}:(-\infty,\infty)\to E^n$ by the formula

$$\hat{x}(t)=\hat{x}+\int_{t_0}^{t}[z(s)+r(s)]\,ds$$

and observe, since for $t>t_0-t_k$

$$x_k(t)=x_k(t_0)+\int_{t_0}^{t}(z_k(s)+G(y_k)(s))\,ds,$$

that $x_k(t)\to\hat{x}(t)$ pointwise in $(-\infty,\infty)$. Furthermore, we observe that as X is closed we have $\hat{x}(t)\in X$ almost everywhere in $(-\infty,\infty)$ and that $\hat{x}$ is locally absolutely continuous. From these facts, and since $s\to g(t-s)$ is Lebesgue integrable on $(-\infty, t)$ we have, by the Dominated Convergence Theorem,

$$\lim_{k\to\infty}\int_{-\infty}^{t} g(t-s)h(x_k(s))\,ds=\int_{-\infty}^{t} g(t-s)h(\hat{x}(s))\,ds.$$

This implies, for all $t\in(-\infty,\infty)$, as a consequence of (4.3), that

$$r(t)=\int_{-\infty}^{t} g(t-s)h(\hat{x}(s))\,ds.$$

It is now easy to see that for all $t\in(-\infty,\infty)$,

$$\begin{aligned}\hat{x}(t)&=\hat{x}+\int_{t_0}^{t}\left(z(s)+\int_{-\infty}^{s} g(s-\tau)h(\hat{x}(\tau))\,d\tau\right)ds\\&=\hat{x}(0)+\int_{0}^{t}\left(z(s)+\int_{-\infty}^{s} g(s-\tau)h(\hat{x}(\tau))\,d\tau\right)ds.\end{aligned} \tag{4.4}$$

We now wish to show that $\hat{x}$ is a trajectory. To this end we appeal to the Lower Closure Theorem 3.2 using the following notation. For each integer $N=\pm1, \pm2, \cdots$, we let

(i) $G=[N, N+1)$, $G_k=[N, N+1)$, $k=1, 2, \cdots$;

(ii) $\eta_k(t)=L(x_k(t), u_k(t))$, $\xi_k(t)=z_k(t)$, $\xi(t)=z(t)$, $\lambda_k(t)\equiv 0$, $\lambda(t)\equiv 0$, $x_k(t)$ as above and $x(t)=\hat{x}(t)$.

Observe that we also have

$$\begin{aligned}\lim_{k\to\infty}\int_{G_k}\eta_k(t)\,dt&=\lim_{k\to\infty}\int_{N}^{N+1} L(x_k(t), u_k(t))\,dt\\&=\lim_{k\to\infty}\int_{N+t_k-t_0}^{N+1+t_k-t_0} L(x(s), u(s))\,ds=0,\end{aligned}$$

since $L(x, u)$ is nonnegative and (4.2) holds. Thus, all the hypotheses of Theorem 3.2 are satisfied and we can conclude that there exists an integrable function $\eta_N : [N, N+1) \to E^1$ such that

$$\hat{x}(t) \in X \quad \text{and} \quad (\eta_N(t), z(t)) \in \tilde{\mathbb{Q}}(\hat{x}(t)), \quad \text{a.e. } t \in [N, N+1)$$

and

$$\int_N^{N+1} \eta_N(t)\, dt \leqq 0.$$

By standard measurable selection arguments (see Cesari [9, Thm. 11.4i]), there exists a measurable function $u_N : [N, N+1) \to E^m$ such that for almost all $t \in [N, N+1)$ we have $\eta_N(t) \geqq L(\hat{x}(t), u_N(t))$, $z(t) = f(\hat{x}(t), u_N(t))$, and $u_N(t) \in U(\hat{x}(t))$. The desired control generating the trajectory $\hat{x} : (-\infty, \infty) \to E^n$ is now obtained by defining $\hat{u} : (-\infty, \infty) \to E^m$ by

$$\hat{u}(t) \equiv u_N(t), \qquad N \leqq t < N+1,$$

$N = \pm 1, \pm 2, \cdots$. Clearly, $\hat{u}$ is measurable and upon substituting into (4.4) we have for any $t \in E^1$,

$$\hat{x}(t) = \hat{x}(0) + \int_0^t f(\hat{x}(s), \hat{u}(s))\, ds + \int_0^t \int_{-\infty}^s g(s-\tau) h(\hat{x}(\tau))\, d\tau,$$

implying that $\hat{x}$ is a trajectory. Moreover, for any integer N, we have

$$0 \leqq \int_N^{N+1} L(\hat{x}(t), \hat{u}(t))\, dt \leqq \int_N^{N+1} \eta_N(t)\, dt = 0,$$

from which it follows that $L(\hat{x}(t), \hat{u}(t)) = 0$ for almost all $t \in E^1$ since L is nonnegative. Thus we see that $\hat{x} \in \mathcal{F}$. This, however, leads to a contradiction since for all k,

$$\varepsilon_0 \leqq |x_k(t_0) - \bar{x}| \leqq |x_k(t_0) - \hat{x}(t_0)| + |\hat{x}(t_0) - \bar{x}| < |x_k(t_0) - \hat{x}(t_0)| + \frac{\varepsilon_0}{2},$$

giving us the contradiction

$$0 = \lim_{k \to \infty} |x_k(t_0) - \hat{x}(t_0)| > \frac{\varepsilon_0}{2}.$$

Therefore we must have $x(t) \to \bar{x}$ as $t \to \infty$.

Before presenting the desired existence results we require the following technical lemma.

LEMMA 4.2. *Let $X \subset E^n$ be closed, let $h : X \to E^p$ be continuous, and let g be an $n \times p$ matrix function satisfying* (2.5). *Suppose that $\{x_i, u_i\} \in A_0$ for $i = 1, 2$ are such that $\lim_{T \to \infty} x_i(T) = \bar{x}$, $i = 1, 2$. Then we have*

$$\lim_{T \to \infty} \int_0^T \left(\int_{T-t}^{\infty} g(s)\, ds \right) [h(x_1(t)) - h(x_2(t))]\, dt = 0.$$

Proof. Let $\varepsilon > 0$ be given. From the continuity of h and the convergence of $x_i(T)$ to $\bar{x}$ as $T \to \infty$, there exists $\tau > 0$ so that for all $T \geqq \tau$,

$$|h(x_1(T)) - h(x_2(T))| < \varepsilon.$$

For $T > \tau$ we write

$$\left|\int_0^T \left(\int_{T-t}^{\infty} g(s)\,ds\right)[h(x_1(t)) - h(x_2(t))]\,dt\right|$$
$$\leqq \left|\int_0^{\tau} \left(\int_{T-t}^{\infty} g(s)\,ds\right)[h(x_1(t)) - h(x_2(t))]\,dt\right|$$
$$+ \left|\int_{\tau}^{T} \left(\int_{T-t}^{\infty} g(s)\,ds\right)[h(x_1(t)) - h(x_2(t))]\,dt\right|$$
$$= I_1(T) + I_2(T),$$

and estimate $I_1(T)$ and $I_2(T)$ separately. As x_i, $i = 1, 2$, are admissible trajectories, they are bounded. Thus, as h is continuous, there exists $H > 0$ so that for all $t \in [0, \infty)$, $|h(x_1(t)) - h(x_2(t))| \leqq H$. This gives us

$$I_1(T) \leqq \int_0^{\tau} \int_{T-t}^{+\infty} |g(s)|\,ds |h(x_1(t)) - h(x_2(t))|\,dt$$
$$\leqq H \int_0^{\tau} \int_{T-t}^{\infty} |g(s)|\,ds\,dt = H\left[\int_{T-\tau}^{T} \int_t^{\infty} |g(s)|\,ds\,dt\right]$$
$$= H\left[\int_0^T \int_t^{+\infty} |g(s)|\,ds\,dt - \int_0^{T-\tau} \int_t^{\infty} |g(s)|\,ds\,dt\right],$$

which tends to zero as $T \to \infty$ since

$$\lim_{T\to\infty} \int_0^T \int_t^{\infty} |g(s)|\,ds\,dt = \int_0^{\infty} \int_t^{\infty} |g(s)|\,ds\,dt = \int_0^{\infty} t|g(t)|\,dt < \infty.$$

For $I_2(T)$ we observe that

$$I_2(T) \leqq \int_{\tau}^{T} \left(\int_{T-t}^{\infty} |g(s)|\,ds\right)|h(x_1(t)) - h(x_2(t))|\,dt < \varepsilon \int_{\tau}^{T} \int_{T-t}^{\infty} |g(s)|\,ds\,dt$$
$$\leqq \varepsilon \int_0^{\infty} \int_t^{\infty} |g(s)|\,ds\,dt,$$

and since $\varepsilon > 0$ was arbitrary, it follows that for T sufficiently large we have that $I_2(T)$ is as small as desired. Combining these two results, we obtain

$$\lim_{T\to\infty} \int_0^T \left(\int_{T-t}^{+\infty} g(s)\,ds\right)[h(x_1(t)) - h(x_2(t))]\,dt = 0.$$

We now present the following set of sufficient conditions for catching-up optimality.

THEOREM 4.1. *Assume that $X \subset E^n$ is closed, $U : X \to 2^{E^m}$ is a set-valued mapping with closed graph M, $f^0 : M \to E^1$ is lower semicontinuous, $f : M \to E^n$ and $h : X \to E^p$ are continuous, and g is an $n \times p$ matrix function satisfying* (2.5). *In addition, assume that* (A1)-(A4) *hold. If $\{x^*, u^*\} \in A_0$ is such that*

$$\text{(4.5)} \qquad \text{(i)} \quad \int_0^{\infty} L(x^*(t), u^*(t))\,dt < \infty,$$

and

$$\text{(4.6)} \qquad \text{(ii)} \quad \int_0^{\infty} L(x^*(t), u^*(t))\,dt \leqq \lim_{T\to\infty} \int_0^T L(x(t), u(t))\,dt,$$

for all $\{x, u\} \in A_0$, *where the limit on the right is either finite or positive infinity, then the pair* $\{x^*, u^*\}$ *is catching-up optimal.*

Proof. Let $\{x^*, u^*\} \in A_0$ be as above and let $\{x, u\} \in A_0$ be arbitrary. Then for any $T > 0$, we have

$$\begin{aligned} J_T(x, u) - J_T(x^*, u^*) &= \int_0^T L(x(t), u(t))\, dt - \int_0^T L(x^*(t), u^*(t))\, dt \\ &\quad + \left\langle \bar{p}, \int_0^T \left[f(x^*(t), u^*(t)) + \left(\int_0^\infty g(s)\, ds \right) h(x^*(t)) \right] dt \right\rangle \\ &\quad - \left\langle \bar{p}, \int_0^T \left[f(x(t), u(t)) + \left(\int_0^\infty g(s)\, ds \right) h(x(t)) \right] dt \right\rangle \\ &= \int_0^T L(x(t), u(t))\, dt - \int_0^T L(x^*(t), u^*(t))\, dt \\ &\quad + \left\langle \bar{p}, \int_0^T \left[f(x^*(t), u^*(t)) + \left(\int_0^{T-t} g(s)\, ds \right) h(x^*(t)) \right] dt \right\rangle \\ &\quad - \left\langle \bar{p}, \int_0^T \left[f(x(t), u(t)) + \left(\int_0^{T-t} g(s)\, ds \right) h(x(t)) \right] dt \right\rangle \\ &\quad + \left\langle \bar{p}, \int_0^T \left(\int_{T-t}^\infty g(s)\, ds \right) [h(x^*(t)) - h(x(t))]\, dt \right\rangle. \end{aligned}$$

Also, for $T > 0$ and $\{x, u\} \in A_0$, we observe that

$$\begin{aligned} \int_0^T \left(\int_0^{T-t} g(s)\, ds \right) h(x(t))\, dt &= \int_0^T \left(\int_t^T g(s-t)\, ds \right) h(x(t))\, dt \\ &= \int_0^T \int_0^t g(t-s) h(x(s))\, ds\, dt, \end{aligned}$$

so that

$$\begin{aligned} J_T(x, u) - J_T(x^*, u^*) &= \int_0^T [L(x(t), u(t)) - L(x^*(t), u^*(t))]\, dt \\ &\quad + \left\langle \bar{p}, \int_0^T (\dot{x}^*(t) - \dot{x}(t))\, dt \right\rangle \\ &\quad + \left\langle \bar{p}, \int_0^T \left(\int_{T-t}^\infty g(s)\, ds \right) [h(x^*(t)) - h(x(t))]\, dt \right\rangle \\ &= \int_0^T [L(x(t), u(t)) - L(x^*(t), u^*(t))]\, dt + \langle \bar{p}, x^*(T) - x(T) \rangle \\ &\quad + \left\langle \bar{p}, \int_0^T \left(\int_{T-t}^\infty g(s)\, ds \right) [h(x^*(t)) - h(x(t))]\, dt \right\rangle. \end{aligned}$$

We now suppose,

$$\int_0^\infty L(x(t), u(t))\, dt < \infty. \tag{4.7}$$

In this case, as a consequence of Proposition 4.2, $\lim_{T \to \infty} x(T) = \bar{x}$ and we obtain

$$\lim_{T \to \infty} [J_T(x, u) - J_T(x^*, u^*)] = \int_0^{+\infty} L(x(t), u(t))\, dt - \int_0^{+\infty} L(x^*(t), u^*(t))\, dt \geqq 0,$$

where we have used the fact that $x^*(T) \to \bar{x}$ as $T \to \infty$ and applied Lemma 4.1. Thus, it is clear that for any $\varepsilon > 0$ we can find $T_\varepsilon = T(\varepsilon, x, u) \geqq 0$ so that for all $T \geqq T_\varepsilon$

$$J_T(x^*, u^*) < J_T(x, u) + \varepsilon.$$

In the case where the improper integral (4.7) is not finite, the boundedness of both x and x^* imply there exist constants M and K such that for all $T > 0$,

$$|\langle \bar{p}, x^*(T) - x(T) \rangle| < M|\bar{p}|$$

and

$$\begin{aligned}
&\left| \left\langle \bar{p}, \int_0^T \left(\int_{T-t}^{+\infty} g(s)\, ds \right) [h(x^*(t)) - h(x(t))]\, dt \right\rangle \right| \\
&\quad \leqq |\bar{p}| K \int_0^T \left(\int_{T-t}^{\infty} |g(s)|\, ds \right) dt \\
&\quad \leqq |\bar{p}| K \int_0^\infty \left(\int_t^\infty |g(s)|\, ds\, dt \right) \\
&\quad = |\bar{p}| K \int_0^\infty t|g(t)|\, dt.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\liminf_{T \to \infty} [J_T(x, u) - J_T(x^*, u^*)] \\
&\quad \geqq \liminf_{T \to \infty} \left\{ \int_0^T [L(x(t), u(t)) - L(x^*(t), u^*(t))]\, dt - |\bar{p}| M - |\bar{p}| K \int_0^\infty t|g(t)|\, dt \right\} \\
&\quad = +\infty.
\end{aligned}$$

Therefore in each case we arrive at the desired result, and so $\{x^*, u^*\}$ is catching-up optimal.

We are now ready to give the main result of this paper. As a result of the previous theorem, it is clear that a catching up optimal solution exists if we can establish the existence of a strongly optimal solution of the associated problem consisting of minimizing

$$I(x, u) = \int_0^\infty L(x(t), u(t))\, dt \tag{4.8}$$

over all $\{x, u\} \in A_0$. With this brief remark we give the following result.

THEOREM 4.2. *Let $X \subset E^n$ be closed, let $U : X \to 2^{E^m}$ have a closed graph M, let $f^0 : X \to E^1$ be lower semicontinuous, let $f : M \to E^n$ and $h : X \to E^p$ be continuous, and let g be an $n \times p$ matrix valued function satisfying* (2.5). *Furthermore, suppose that* (A1)-(A4) *hold and that there exists an admissible pair $\{\hat{x}, \hat{u}\} \in A_0$ such that* (4.8) *is finite. Under these conditions, the optimal control problem described by* (2.1)-(2.4) *and* (2.6) *has a catching-up optimal solution.*

Proof. By hypothesis, we have

$$0 \leqq \inf_{\{x,u\} \in A} I(x, u) \leqq I(\hat{x}, \hat{u}) < \infty.$$

Thus there exists a minimizing sequence $\{x_k, u_k\}_{k=1}^{\infty}$ for the associated optimal control problem. As a consequence of Proposition 2.1, f, h, and L satisfy the growth condition (A2) with

$$\int_0^T L(x_k(t), u_k(t))\, dt \leqq \int_0^{+\infty} L(\hat{x}(t), \hat{u}(t))\, dt < \infty$$

for all $T > 0$. Therefore, by a standard diagonalization process, there exists a subsequence, say still $\{x_k, u_k\}$, and locally integrable functions $z:[0,\infty) \to E^n$ and $y:[0,\infty) \to E^p$ such that $f(x_k(t), u_k(t)) \to z(t)$ and $h(x_k(t)) \to y(t)$ weakly in $L^1_{\text{loc}}([0,\infty); E^n)$ and $L^1_{\text{loc}}((0,\infty); E^p)$, respectively. This implies that for each $t \geqq 0$ we have

$$\begin{aligned} x_k(t) &= x_0(0) + \int_0^t f(x_k(s), u_k(s))\, ds + \int_0^t \left(\int_{-\infty}^0 g(s-\tau) h(x_0(\tau))\, d\tau \right) dt \\ &\quad + \int_0^t \int_0^s g(s-\tau) h(x_k(\tau))\, d\tau\, dt \\ &= x_0(0) + \int_0^t f(x_k(s), u_k(s))\, ds + \int_0^t \left(\int_{-\infty}^0 g(s-\tau) h(x_0(\tau))\, d\tau \right) ds \\ &\quad + \int_0^t \left(\int_\tau^t g(s-\tau)\, ds \right) h(x_k(\tau))\, d\tau, \end{aligned}$$

which clearly converges pointwise to $x^*:[0,\infty) \to E^n$ given by

$$\begin{aligned} x^*(t) &= x_0(0) + \int_0^t z(s)\, ds + \int_0^t \left(\int_{-\infty}^0 g(s-\tau) h(x_0(\tau))\, d\tau \right) ds \\ &\quad + \int_0^t \left(\int_\tau^t g(s-\tau)\, ds \right) y(\tau)\, d\tau \\ &= x_0(0) + \int_0^t z(s)\, ds + \int_0^t \left(\int_{-\infty}^0 g(s-\tau) h(x_0(\tau))\, d\tau \right) ds \\ &\quad + \int_0^t \left(\int_0^s g(s-\tau) y(\tau)\, d\tau \right) ds. \end{aligned} \tag{4.9}$$

Moreover, this pointwise convergence gives us

$$\lim_{k \to \infty} h(x_k(t)) = h(x^*(t)),$$

so that we have $y(t) = h(x^*(t))$ almost everywhere $t \in E^1$. We now apply the Lower Closure Theorem 3.1 using the following notation:

(i) $G = [0,\infty)$, $G_k = [0, k]$,

(ii) $\eta_k(t) = L(x_k(t), u_k(t))$, $\xi_k(t) = f(x_k(t), u_k(t))$, $x_k(t)$ as above and $\lambda_k(t) \equiv 0$, almost everywhere $t \in G_k$, $k = 1, 2, \cdots$,

(iii) $\lambda(t) \equiv 0$, $\xi(t) = z(t)$, and $x(t) = x^*(t)$ almost everywhere $t \geqq 0$.

(iv) $R(x) = \tilde{\mathbb{Q}}_L(x)$.

It is easy to establish that the hypotheses of Theorem 3.1 are indeed met so that we conclude that there exists $\eta: [0,\infty) \to E^1$, which is integrable and satisfies

$$(\eta(t), \xi(t)) \in \tilde{\mathbb{Q}}_L(x(t)) \quad \text{and} \quad x^*(t) \in X \quad \text{a.e. } t \geqq 0$$

and

$$0 \leqq \int_0^\infty \eta(t)\, dt \leqq \liminf_{k \to \infty} \int_0^k L(x_k(t), u_k(t))\, dt \leqq \inf_{\{x,u\} \in A_0} I(x, u).$$

By appealing to standard measurable selection arguments (see, e.g., Cesari [9; Thm. 11.4i]) there exists a measurable function $u^*:[0,\infty)\to E^m$ such that $\eta(t)\geqq L(x^*(t), u^*(t))$, $z(t)=f(x^*(t), u^*(t))$, and $u^*(t)\in U(x^*(t))$ a.e. $0\leqq t$. Substituting this information into (4.8), we obtain

$$x^*(t)=x_0(0)+\int_0^t f(x^*(s), u^*(s))\,ds+\int_0^t\left(\int_{-\infty}^0 g(s-\tau)h(x_0(\tau))\,d\tau\right)ds$$
$$+\int_0^t\left(\int_0^s g(s-\tau)h(x^*(\tau))\,d\tau\right)ds,$$

which clearly shows the pair $\{x^*, u^*\}$ is an admissible pair (here we have extended x^* to $(-\infty, 0]$ by defining $x^*(s)=x_0(s)$). In addition, we further observe that

$$0\leqq\int_0^{+\infty} L(x^*(t), u^*(t))\,dt\leqq\int_0^{+\infty}\eta(t)\,dt\leqq\inf_{\{x,u\}\in A_0} I(x, u),$$

which shows that $\{x^*, u^*\}$ minimizes $I(x, u)$ over A_0. The desired conclusion now follows by a direct application of Theorem 4.1.

The above existence result generalizes the works of Brock and Haurie [5] and Leizarowitz [15] in two directions. The first of these is the obvious extension to models exhibiting time delay in the state variable. As regards this extension, we also note that our formulation also implies results for a class of problems with finite lag where for some fixed $r>0$ we assume $g(s)\equiv 0$. In such a situation, the hypothesis (2.5) concerning the kernel g are valid if g is essentially bounded on $[0, r]$. The other direction in which we have generalized these earlier results is that we have weakened both the convexity and growth hypotheses. Indeed, in [5] and [15], it is required that the set

$$\Omega=\{(x, z^0, z): x\in X, z^0\geqq f^0(x, u), z=f(x, u), u\in U(x)\}$$

is closed and convex. This condition is stronger than the convexity and upper semicontinuity conditions we require of the sets $\tilde{Q}(x)$. On the other hand, to ensure that assumption (A4′), and hence (A4) hold, we must, in general, assume these stronger convexity hypotheses. However, they need not always be assumed, as is seen in Example 4.1. As regards the growth condition in [15] we have already observed that our growth condition is weaker (see Remark 2.2).

The most difficult hypotheses to be satisfied in the above work concern our condition (A4) and the existence of an admissible pair $\{\hat{x}, \hat{u}\}$ for which $I(\hat{x}, \hat{u})$ is finite. For conditions that ensure (A4) holds we refer the reader to Proposition 4.1. For the existence of $\{\hat{x}, \hat{u}\}$ we note that in the case of finite delay, this condition can be realized by controlling from the initial function $x_0(s)$, $-r\leqq s\leqq 0$, to the terminal function $\bar{x}$, $-r\leqq s\leqq 0$ in finite time. Since the terminal function is a constant function, this problem can be addressed by utilizing known null controllability results for problems with time delay (see, e.g., [1], [10], [11], [19]). For the case of infinite delay such an approach is not applicable and the realization of this hypothesis requires further investigation.

5. Examples.

Example 5.1. In this example we return once more to Example 4.1. As we have already seen, this examples satisfies both assumptions (A3) and (A4). Also as $(x, u)\in[-1, 1]\times[-1, 1]$, a compact set, it is easy to see that the growth condition (A2) is also satisfied. Furthermore, the linearity of $f^0(x, u)=2x(4x^2+6x-9)-9u$ and $f(x, u)=u+2x$ with respect to u easily ensures that the convexity and seminormality hypotheses

of (A1) are also met. Consequently, to apply Theorem 4.2 it is sufficient to ensure that there exists an admissible pair $\{\hat{x}, \hat{u}\}$ for which

$$0 \leqq \int_0^\infty L(\hat{x}(t), \hat{u}(t))\, dt = \int_0^\infty (4x(t)^2(2x(t)+3))\, dt.$$

In this simple example we note that by choosing the control $\hat{u}(t) \equiv 0$, we have that $\hat{x}(t) = \varphi(0)\, e^{-2t}$ is an admissible trajectory for every continuous function $\varphi: (-\infty, 0] \to [-1, 1]$. Clearly, this admissible pair satisfies the desired controllability property. Thus there exists a catching-up optimal solution for this problem.

Example 5.2. The Ramsey model with delay. In this example we present a generalization of the classical Ramsey model of economic growth. Since we deal with minimization problems we use a slightly nonstandard formulation.

For this model we let $x = x(t)$, $-\infty < t < \infty$, denote the stock of capital at time t and $u = u(t)$, $0 \leqq t$, the consumption. We denote the production function by $h = h(x)$ and the utility by $-f^0 = -f^0(u)$. As is standard practice, we assume that $h(0) = 0$ and that h is a smooth increasing strictly concave function satisfying

$$\lim_{x \to 0^+} h'(x) = \infty; \qquad \lim_{x \to \infty} h'(x) = 0,$$

and furthermore, that f^0 is smooth, decreasing, strictly convex with

$$\lim_{u \to 0^+} f^{0\prime}(u) = -\infty \quad \text{and} \quad \lim_{u \to \infty} f^{0\prime}(u) = 0.$$

To introduce the delay we let $g: [0, \infty) \to [0, 1]$ satisfy the hypotheses outlined in (2.5) and assume

$$\int_0^\infty g(s)\, ds = 1. \tag{5.1}$$

This function is introduced to reflect the reduction in production of a plant or factory due to aging equipment. Thus the term $g(t-s)h(x(s))$ represents the rate of production at time t generated by a stock of capital $x(s)$, s time units ago. With this notation we let $\lambda > 0$ be given and consider the optimal control problem

$$\text{minimize} \int_0^\infty f^0(u(t))\, dt \tag{5.2}$$

subject to

$$\dot{x}(t) = \int_{-\infty}^t g(t-s)h(x(s))\, ds - \lambda x(t) - u(t), \tag{5.3}$$

$$x(s) = x_0(s), \qquad s \leqq 0, \tag{5.4}$$

$$0 \leqq x(t) \leqq \hat{x}, \tag{5.5}$$

$$0 \leqq u(t) \leqq h(x(t)). \tag{5.6}$$

The upper bound $\hat{x} > 0$ given in (5.4) is a standard hypothesis in the nondelay case with $\hat{x}$ chosen to be the unique positive solution to

$$h(x) = \lambda x.$$

We remark that in the nondelay case any admissible stock $x(t) > \hat{x}$ is necessarily decreasing. Thus it is reasonable to ask that $x(t) \leqq \hat{x}$, whenever $x_0(s) \in [0, \hat{x}]$; $s \leqq 0$.

As a result of the boundedness of the admissible states it is an easy matter to see that the growth condition (A2) is satisfied. Moreover, the convexity conditions placed on f^0 and the linearity of $f(x, u) = -\lambda x - u$ ensure that the sets $\tilde{\mathbb{Q}}(x)$ are convex and

closed. Furthermore, as f^0 is continuous, these sets also enjoy property (K) so that (A1) is satisfied.

For the optimal steady state we recall that in the nondelay case (see, e.g., Samuelson [18] or Cass [8]) there exists a unique pair $(\bar{x}, \bar{u})$ such that for all (x, u), $x \geqq 0$, $u \geqq 0$ we have

$$L(x, u) = f^0(u) - f^0(\bar{u}) + f^{0\prime}(\bar{u})[h(x) - \lambda x - u] \geqq 0. \tag{5.7}$$

The pair $(\bar{x}, \bar{u})$ is uniquely determined by the system of equations

$$h'(\bar{x}) = \lambda \quad \text{and} \quad \bar{u} = h(\bar{x}) - \lambda \bar{x}.$$

Moreover, the strict concavity of h and the convexity of f^0 imply that for every $\varepsilon > 0$ there exists a $\delta > 0$ so that if $|x - \bar{x}| > \varepsilon$, then

$$L(x, u) > \delta \quad \text{for all } u \in [0, h(x)]. \tag{5.8}$$

As (5.1) holds, it is now easy to see that (A3) is satisfied with L given by (5.7). Finally, (5.8) is simply the verification of (A4′). Thus all the hypotheses, except for the existence of $\{\hat{x}, \hat{u}\}$ admissible so that $I(\hat{x}, \hat{u}) < \infty$, needed in the previous sections are met and so the existence of a catching-up optimal solution is assured whenever $\{\hat{x}, \hat{u}\}$ exists. As remarked earlier, in the nondelay case, or in the finite-delay case, this hypothesis can be met through the application of known controllability results (in particular, see Angell [1; Thm. 5.2]).

Example 5.3. Optimal exploitation of a renewable resource. We conclude our examples with an economic growth model that incorporates a renewable resource into the classical Ramsey model. As in the previous example we let $x = x(t)$, $t \in \mathbb{R}$, denote the stock of capital at time t and $u(t)$, $t \geqq 0$, denote consumer consumption. The production of capital stock depends on a renewable resource (e.g., a forest) and we let $y(t)$, $t \in \mathbb{R}$, denote the amount of the resource available at time t. This resource is harvested at a rate $v(t)$, $t \geqq 0$. With these variables we introduce the following control system:

$$\dot{x}(t) = f(x(t), v(t)) - \lambda x(t) - u(t), \tag{5.9}$$

$$\dot{y}(t) = -v(t) + \int_{-\infty}^{t} g(t-s)h(x(s), y(s))\, ds \quad \text{a.e. } t \geqq 0, \tag{5.10}$$

$$\begin{pmatrix} x(s) \\ y(s) \end{pmatrix} = \begin{pmatrix} x_0(s) \\ y_0(s) \end{pmatrix} \quad \text{for all } s \in (-\infty, 0], \tag{5.11}$$

$$\begin{aligned} 0 \leqq x(t) &\leqq \hat{x} \\ 0 \leqq y(s) &\leqq \hat{y} \end{aligned} \quad \text{for all } t \leqq 0, \tag{5.12}$$

$$\begin{aligned} 0 \leqq u(t) &\leqq f(x(t), v(t)) \\ 0 \leqq v(t) &\leqq h(x(t), y(t)) \end{aligned} \quad \text{a.e. } 0 \leqq t. \tag{5.13}$$

Here we are letting $f: [0, \infty) \times [0, \infty) \to \mathbb{R}$ denote the production function for the capital stock x be a smooth concave function that satisfies

$$\lim_{x \to 0^+} \frac{\partial f}{\partial x}(x, v) = +\infty \quad \text{for each } v \geqq 0,$$

$$\lim_{x \to \infty} \frac{\partial f}{\partial x}(x, v) = 0 \quad \text{for each } v \geqq 0,$$

$$\frac{\partial f}{\partial x}(x, v) > 0 \quad \text{and} \quad \frac{\partial f}{\partial v}(x, v) > 0 \quad \text{for all } (x, v),$$

$$f(0, v) = 0 \quad \text{for each } v \geqq 0.$$

The function $h: [0, \infty) \times [0, \infty) \to \mathbb{R}$ denotes the rate of growth of resource y and we assume that h is concave and satisfies

$$h(x, 0) = 0 \quad \text{for all } x \geqq 0, \qquad \frac{\partial h}{\partial x}(x, y) \geqq 0 \quad \text{for all } (x, y),$$

and

for each $x \in [0, \infty)$ there exists $y_x > 0$ so that on $(0, y_x)$, $h(x, y) > 0$, $h(x, y_x) = 0$ and on (y_x, ∞), $h(x, y) < 0$.

The function $g: [0, \infty) \to [0, 1]$ satisfies the hypotheses outlined in (2.5) and in addition satisfies

$$\int_0^{+\infty} g(s)\, ds = 1.$$

This function is introduced to reflect the fact that the rate of growth of the resource is age-dependent. Finally, the constant $\lambda > 0$ denotes the rate of depreciation of capital, with x_0 and y_0 given initial functions. We observe that the producer influences the rate of growth of the resource through the level of capital stock as well as through the harvest rate v.

As in the previous example, our objective is to maximize the accumulated utility described by a utility function. $-f^0 = -f^0(u)$ where $f^0: [0, \infty) \to \mathbb{R}$ satisfies the same hypotheses as in Example 5.1. Therefore we consider the optimal control problem

$$\text{minimize} \int_0^{\infty} f^0(u(t))\, dt \tag{5.14}$$

over all pairs of functions $\{(x, y), (u, v)\}$ satisfying (5.9)–(5.13).

The associated optimal steady-state problem becomes

$$\text{minimize } f^0(u)$$

over all $(x, y, u, v) \in M$ satisfying

$$0 = f(x, v) - \lambda x - u, \qquad 0 = h(x, y) - v,$$

where

$$M = \{(x, y, u, v): 0 \leqq x \leqq \hat{x}, 0 \leqq y \leqq \hat{y}, 0 \leqq u \leqq f(x, v), 0 \leqq v \leqq h(x, y)\}.$$

It is easy to see that, under the hypotheses placed on the model, the conditions of Proposition 4.1 are satisfied. Indeed we easily see that M is compact convex and int $M \neq \varnothing$, that $F_1(x, y, u, v) = f(x, v) - \lambda x - u$ and $F_2(x, y, u, v) = h(x, y) - v$ are concave, there exists (x, y, u, v) so that $F_i(x, y, u, v) > 0$, $i = 1, 2$, and that $f^0(\cdot)$ is strictly convex. To establish the remaining hypothesis we observe that if $(x, y, u, v) \in M$ is such that $F_i(x, y, u, v) \geqq 0$ for $i = 1, 2$ we may take $(\hat{u}, \hat{v})$ to be given by

$$\hat{v} = h(x, y) \quad \text{and} \quad \hat{u} = f(x, \hat{v}) - \lambda x,$$

and observe that $\hat{v} \geqq v$ so that $f(x, \hat{v}) \geqq f(x, v)$, which implies

$$u \leqq f(x, v) - \lambda x \leqq f(x, \hat{v}) - \lambda x = \hat{u} \leqq f(x, \hat{v}).$$

Clearly, this implies that $(x, y, \hat{u}, \hat{v}) \in M$ and that $f^0(\hat{u}) \leqq f^0(u)$ as observed. Therefore the only remaining hypothesis in Theorem 4.2 concerns the existence of an admissible

pair $\{(\hat{x}, \hat{y}), (\hat{u}, \hat{v})\}$ for which the improper integral $I[(\hat{x}, \hat{y}, \hat{u}, \hat{v})]$ given by (4.7) is finite. Here, as remarked previously, in the case of finite delay, this hypothesis may be met through the application of known controllability results.

Acknowledgment. The author thanks an anonymous referee for several suggestions and comments that enhanced the presentation of the above results.

REFERENCES

[1] T. S. ANGELL, *On controllability for non-linear hereditary systems: a fixed point approach*, Nonlinear Anal., 4 (1980), pp. 529-545.

[2] E. J. BALDER, *An existence result for optimal economic growth problems*, J. Math. Anal. Appl., 95 (1983), pp. 195-213.

[3] G. R. BATES, *Lower closure and existence theorems for optimal control problems with infinite horizon*, J. Optim. Theory Appl., 24 (1978), pp. 639-649.

[4] R. F. BAUM, *Existence theorems for Lagrange control problems with unbounded time domain*, J. Optim. Theory Appl., 19 (1976), pp. 89-116.

[5] W. BROCK AND A. HAURIE, *On existence of overtaking optimal trajectories over an infinite time horizon*, Math. Oper. Res., 1 (1976), pp. 337-346.

[6] D. A. CARLSON AND A. HAURIE, *Infinite horizon optimal control: theory and applications*, Lecture Notes in Economics and Mathematical Systems, 290, Springer-Verlag, New York, 1987.

[7] D. A. CARLSON, A. HAURIE, AND A. JABRANE, *Existence of overtaking optimal solutions to infinite dimensional control problems on unbounded time intervals*, SIAM J. Control Optim., 25 (1987), pp. 1517-1541.

[8] D. CASS, *Optimum growth in an aggregative model of capital accumulation: a turnpike theorem*, Econometrica, 34 (1966), pp. 833-850.

[9] L. CESARI, *Optimization-Theory and Applications*, Springer-Verlag, New York, 1983.

[10] E. N. CHUKWU, *Null controllability in function space of nonlinear retarded systems with limited control*, J. Math. Anal. Appl., 103 (1984), pp. 198-210.

[11] ———, *On the null-controllability of nonlinear delay systems with restrained controls*, J. Math. Anal. Appl., 76 (1980), pp. 283-296.

[12] R. F. HARTL, *Optimal dynamic advertising policies for hereditary processes*, J. Optim. Theory Appl., 43 (1984), pp. 51-72.

[13] R. F. HARTL AND S. P. SETHI, *Optimal control of a class of systems with continuous lags: dynamic programming approach and economic interpretations*, J. Optim. Theory Appl., 43 (1984), pp. 73-88.

[14] M. KALECKI, *A macrodynamic theory of business cycles*, Econometrica, 3 (1935), pp. 327-344.

[15] A. LEIZAROWITZ, *Existence of overtaking optimal trajectories for problems with convex integrands*, Math. Oper. Res., 10 (1985), pp. 450-461.

[16] W. W. LEONTIEF, *Lags and stability of dynamic systems*, Econometrica, 29 (1961), pp. 659-669.

[17] M. MARCUS AND V. J. MIZEL, *Limiting equations for problems involving long range memory*, Mem. Amer. Math. Soc., 43 (1983).

[18] P. A. SAMUELSON, *A caternary turnpike theorem involving consumption and the golden rule*, Amer. Econ. Rev., 55 (1965), pp. 486-496.

[19] R. G. UNDERWOOD AND D. YOUNG, *Null controllability of nonlinear functional differential equations*, SIAM J. Control Optim., 17 (1979), pp. 753-772.

[20] C. C. VON WEIZSÄCKER, *Existence of optimal programs of accumulation for an infinite time horizon*, Rev. Econom. Stud., 32 (1965), pp. 85-103.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 423-437, March 1990

010

STABILIZATION OF BEAMS BY POINTWISE FEEDBACK CONTROL*

F. CONRAD†

Abstract. A flexible structure consisting of serially connected Euler-Bernoulli beams with co-located sensors and actuators is considered. Controls are point forces and point bending moments applied at the nodes. It is known that uniform exponential stability can be achieved with linear velocity feedback. A sensitivity analysis of the system's spectrum with respect to feedback coefficients is set up. It is also proved that in a particular case exponential decay rate can be obtained from the spectrum of the system.

Key words. serially connected beams, point control, exponential stabilization

AMS(MOS) subject classifications. 93D15, 73K12, 35P10

1. Introduction. We study the transverse deflection of a system of N serially connected Euler-Bernoulli beams as shown in Fig. 1, where each beam has mass density m_i and flexural rigidity E_iI_i.

The left end $x=0$ is clamped. At the right end and at interior nodes, point control forces U_{0i} and point control bending moments U_{1i} are applied. We assume that at each node, the controls are linear combinations of the transverse and angular velocities; thus sensors and actuators are co-located.

This system has been studied extensively by Chen et al. (see, for instance, [2] where major results can be found). For convenience, we summarize the material of [2], which is useful to us.

Let $y(x, t)$ denote the transverse deflection of the structure at time t. In the sequel, we set $\dot{y}(x, t)=\partial y/\partial t(x, t)$ and $y'(x, t)=\partial y/\partial x(x, t)$.

For the open-loop system we get the following equations:

$$
\begin{aligned}
&m_i\ddot{y}+E_iI_iy''''=0, \quad x_{i-1}<x<x_i, \quad i=1, N,\\
&y(0, t)=y'(0, t)=0,\\
&y(x_i^-, t)=y(x_i^+, t), \qquad i=1, N-1,\\
&y'(x_i^-, t)=y'(x_i^+, t), \qquad i=1, N-1,\\
&E_iI_iy'''(x_i^-, t)-E_{i+1}I_{i+1}y'''(x_i^+, t)=U_{0i}, \qquad i=1, N-1,\\
&-E_iI_iy''(x_i^-, t)+E_{i+1}I_{i+1}y''(x_i^+, t)=U_{1i}, \qquad i=1, N-1,\\
&E_NI_Ny'''(L, t)=U_{0N},\\
&-E_NI_Ny''(L, t)=U_{1N}.
\end{aligned}
\tag{1.1}
$$

For the closed-loop system, we add the feedback law

$$
\begin{pmatrix} U_{0i}(t) \\ U_{1i}(t) \end{pmatrix} = K_i \begin{pmatrix} \dot{y}(x_i, t) \\ \dot{y}'(x_i, t) \end{pmatrix}, \qquad i=1, N \tag{1.2}
$$

where K_i is a 2×2 real matrix.

This system can be written in an abstract form [2]. Let $H=L^2(0, L)$, let $V=\{v\in H^2(0, L)\mid v(0)=v'(0)=0\}$ be the energy space, and let $A\in\mathcal{L}(V, V')$ be defined by

$$
\langle Av, w\rangle=\sum_{i=1}^{N} E_iI_i\int_{x_{i-1}}^{x_i} v''(x)w''(x)\,dx \quad \forall v\in V, \quad w\in V.
$$

* Received by the editors May 31, 1988; accepted for publication (in revised form) May 26, 1989.

† Université de Nancy I, UA CNRS 750, B.P. 239, 54506-Vandoeuvre lès Nancy, France. The major part of this work was done while the author was visiting the CMA, Ecole des Mines, at Sophia Antipolis.

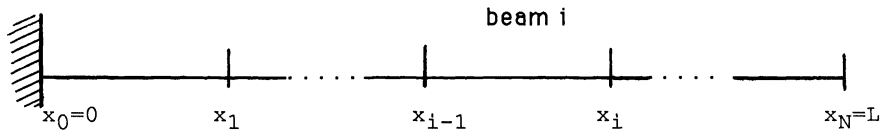


FIG. 1

The observation operator $C \in \mathscr{L}(V, \mathbb{R}^{2N})$ is defined by

$$Cv = (v(x_1), v'(x_1), \cdots, v(x_i), v'(x_i), \cdots, v(x_N), v'(x_N)).$$

The control operator is the transpose $C^* \in \mathscr{L}(\mathbb{R}^{2N}, V')$ of the observation operator.

Let K be the $2N \times 2N$ block-diagonal matrix with entries K_i, $i=1, N$, and M denote the isomorphism $f \to Mf$ on $L^2(0, L)$ where $M(x)$ is the function $M(x) = \sum_{i=1}^{N} m_i 1_{(x_{i-1}, x_i)}(x)$.

In this framework, the equations for the closed-loop system are

$$M\ddot{y} + Ay + C^*KC\dot{y} = 0 \quad \text{in } V'; \quad y \in V. \tag{1.3}$$

THEOREM 1.1 [2]. *Assume K is coercive. Given $(y_0, y_1) \in V \oplus H$, (1.3) admits a unique solution $y \in \mathscr{C}(0, T; V)$; $\dot{y} \in \mathscr{C}(0, T; H)$; $C\dot{y} \in L^2(0, T; \mathbb{R}^{2N})$; $M\ddot{y} \in L^2(0, T; V')$.*

In fact (1.3) defines a $\mathscr{C}^0$ semigroup of contractions on $V \oplus H$ that will be denoted by S_K.

The generator of S_K is

$$\mathscr{A}(K) = \begin{pmatrix} 0 & I \\ -M^{-1}A & -M^{-1}C^*KC \end{pmatrix}.$$

Remark. For a sharp existence result for the open-loop system, we refer the reader to [9].

The main result in [2] concerns stabilization. The elastic energy of the system is defined as

$$E(t) = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} [m_i \dot{y}^2(x, t) + E_i I_i y''^2(x, t)]\, dx.$$

THEOREM 1.2 [2]. *Assume the following.*

(i) $m_i \leqq m_{i+1}$ *and* $E_i I_i \geqq E_{i+1} I_{i+1}$, $i = 1, N-1$.

(ii) $K_i = \left(\begin{smallmatrix} \alpha_i & 0 \\ 0 & 0 \end{smallmatrix}\right)$ *with* $\alpha_i \geqq 0$, $i = 1, N-1$.

(iii) *Either* $K_N = \left(\begin{smallmatrix} \alpha_N & 0 \\ 0 & \beta_N \end{smallmatrix}\right)$ *with* $\alpha_N > 0$, $\beta_N \geqq 0$, *or*

$$K_N = \begin{pmatrix} \alpha_N & -\gamma_N \\ \gamma_N & \beta_N \end{pmatrix}, \qquad \alpha_N > 0, \quad \beta_N > 0, \quad \gamma_N \in \mathbb{R}.$$

Then uniform exponential stability holds: there exists $\mu > 0$ such that $E(t) \leqq$ const. $e^{-\mu t} E(0)$, for all $(y_0, y_1) \in V \oplus H$ or else

$$\|S_K(t)\|_{V \oplus H} \leqq \text{const. } e^{-\mu t}. \tag{1.4}$$

Remark 1. Assumptions (ii) and (iii) do not imply that K is coercive. However, the conclusions of Theorem 1.1 are valid provided the estimate on $C\dot{y}$ is replaced by $\dot{y}(x_i, \cdot) \in L^2(0, T)$ whenever $\alpha_i > 0$, $i = 1, \cdots, N$, and $\dot{y}'(x_N, \cdot) \in L^2(0, T)$ whenever $\beta_N > 0$.

Remark 2. Exponential decay for an Euler-Bernoulli beam with rate control on the bending moment only has been obtained in [3] by means of a frequency domain technique. The energy multiplier method used to prove Theorem 1.2 does not seem to work in that case.

The optimal rate of decay μ in (1.4) is of course the spectral radius ω of the semigroup S_K: $\omega = \lim t^{-1} \log \|S_K(t)\|_{V\oplus H}$ as $t \to \infty$.

It was conjectured in [2] that the rate of decay of the energy is given by $\rho(K) = \sup\{\operatorname{Re} \nu_p / p \in \mathbb{N}\}$, where the ν_p are the eigenvalues of the generator $\mathscr{A}(K)$. This cannot result from standard theorems, since, due to the existence of a vertical asymptote for the ν_p in the complex plane (see [3]-[5] and [10]), S_K is neither differentiable nor compact. We prove the conjecture in a particular case: *for a single homogeneous beam with rate control on the shear only, $\rho(K) = \sup\{\operatorname{Re} \nu_p / p \in \mathbb{N}\}$, at least for small K (Corollary 4.5 below).*

Assuming the conjecture holds, the spectrum of the system for a single homogeneous beam has been computed in [2] with respect to feedback coefficients, leading to interesting conclusions. In this paper, *we obtain analytical results concerning the dependence of the spectrum with respect to K, for a general system of connected beams (Theorem 2.2 below).* Applied to a single beam, the results reinforce the conclusions obtained in [2].

Our results are obtained by the following perturbation technique. We consider $\mathscr{A}(K)$ as a perturbation of $\mathscr{A}(0)$ (the free vibrating system) and then get expansions of the eigenvalues and eigenvectors. Therefore, due to the technique, the results, although very precise, are only valid for small K.

The summary of the paper is as follows. In § 2, we set up a sensitivity analysis to get estimates for the eigenvalues of $\mathscr{A}(K)$, for N connected beams. In § 3, we apply the result to a single beam. In § 4, also for a single beam, we improve the estimates on the spectrum, get estimates for the eigenvectors, and then prove the conjecture mentioned previously.

2. Sensitivity analysis of the spectrum of the system. We recall the generator of the semigroup $S_K(t)$:

$$\mathscr{A}(K) = \begin{pmatrix} 0 & I \\ -M^{-1}A & -M^{-1}C^*KC \end{pmatrix}.$$

Let W be the subspace of functions of V that are piecewise H^4 on each interval (x_{i-1}, x_i). Then the domain of the unbounded operator $\mathscr{A}(K)$ on $V \oplus H$ is $\operatorname{dom} \mathscr{A}(K) = \{(u, v) \mid u \in W,\ v \in V;\ Bu = KCv\}$ where

$$Bu = \{\cdots E_i I_i u'''(x_i^-) - E_{i+1} E_{i+1} u'''(x_i^+), -E_i I_i u''(x_i^-) + E_{i+1} I_{i+1} u''(x_i^+), \cdots, E_N I_N u'''(L), -E_N I_N u''(L)\}.$$

The spectrum of $\mathscr{A}(K)$ is purely point spectrum, by classical regularity results (see also [10]).

Obviously, $\nu \in \mathbb{C}$ is an eigenvalue, $\binom{\phi}{\psi} \in \operatorname{dom} \mathscr{A}(K)$ is an eigenvector of $\mathscr{A}(K)$ if and only if

$$\psi = \nu\phi, \tag{2.1}$$

$$A\phi + \nu^2 M\phi + \nu C^* KC\phi = 0. \tag{2.2}$$

Therefore, we only must solve (2.2). This will be achieved by means of the Implicit Function Theorem in an adequate framework, to get the eigenelement as functions of K, near $K = 0$. We now develop the methodology.

2.1. Free vibrations of the structure. In case $K = 0$, (2.2) reduces to

$$A\phi + \nu^2 M\phi = 0, \quad \phi \in W, \qquad B\phi = 0, \tag{2.3}$$

or in strong form, with $\phi \in W$:

$$
\begin{aligned}
&E_iI_i\phi'''' + \nu^2 m_i\phi = 0, \quad x_{i-1} < x < x_i, \quad i = 1, N,\\
&E_iI_i\phi'''(x_i^-) - E_{i+1}I_{i+1}\phi'''(x_i^+) = 0, \qquad i = 1, N-1,\\
&-E_iI_i\phi''(x_i^-) + E_{i+1}I_{i+1}\phi''(x_i^+) = 0, \qquad i = 1, N-1,\\
&E_NI_N\phi'''(L) = 0,\\
&-E_NI_N\phi''(L) = 0.
\end{aligned} \tag{2.4}
$$

As an unbounded operator on H, A is self-adjoint and A^{-1} is compact. Let $(\omega_p^2)_{p\in N}$ be the eigenvalues of $A(\omega_p > 0)$ and ϕ_p the associated eigenfunctions. We assume that the eigenvalues of A are geometrically simple. Then (2.3) has a sequence of solutions (ν_p, ϕ_p) with $\nu_p = \pm i\omega_p$.

2.2. Forced vibrations of the structure. In case $K \neq 0$, we consider solutions (ν, ϕ) of (2.2) near (ν_p, ϕ_p), for any fixed p. Since (2.2) is not well-posed for the eigenfunctions, we must normalize ϕ. Define

$\mathcal{W} = \{\phi_p\}^\perp$ in W, for the H-scalar product, and set $\nu = \nu_p + \mu$, $\mu \in \mathbb{C}$, $\phi = \phi_p + w$, $w \in \mathcal{W}$.

Injecting these expansions into (2.2), we get

$$A w + \nu_p^2 M w + (2\nu_p\mu + \mu^2) M(\phi_p + w) + (\nu_p + \mu) C^* K C(\phi_p + w) = 0$$

or in strong form

$$
\begin{aligned}
&E_iI_iw'''' + \nu_p^2 m_i w + (2\nu_p\mu + \mu^2) m_i(\phi_p + w) = 0, \quad x_{i-1} < x < x_i, \quad i = 1, N,\\
&\begin{bmatrix} E_iI_iw'''(x_i^-) - E_{i+1}I_{i+1}w'''(x_i^+) \\ -E_iI_iw''(x_i^-) + E_{i+1}I_{i+1}w''(x_i^+) \end{bmatrix}\\
&\qquad - (\nu_p + \mu) K_i \begin{pmatrix} \phi_p + w \\ \phi_p' + w' \end{pmatrix}(x_i) = 0, \qquad i = 1, N-1,\\
&\begin{bmatrix} E_NI_Nw'''(L) \\ -E_NI_Nw''(L) \end{bmatrix} - (\nu_p + \mu) K_N \begin{pmatrix} \phi_p + w \\ \phi_p' + w' \end{pmatrix}(L) = 0,
\end{aligned} \tag{2.5}
$$

that is, $\mathcal{F}(\mu, w, K) = 0$ where $\mathcal{F}: \mathbb{C} \times \mathcal{W} \times \mathcal{M}_{2N} \to H \times \mathbb{C}^{2N}$ is the left-hand side of (2.5).

Obviously, $\mathcal{F}(0, 0, 0) = 0$ and $\mathcal{F}$ is regular. We want to apply the Implicit Function Theorem to (2.5). Let $\hat{\mu}$, $\hat{w} \in \mathbb{C} \times \mathcal{W}$. Then

$$
\mathcal{F}_{\mu w}(0,0,0)\hat{\mu}\hat{w} = \begin{cases}
E_iI_i\hat{w}'''' + \nu_p^2 m_i\hat{w} + 2\nu_p\hat{\mu}m_i\phi_p, & x_{i-1} < x < x_i, \quad i = 1, N,\\
E_iI_i\hat{w}'''(x_i^-) - E_{i+1}I_{i+1}\hat{w}'''(x_i^+), & i = 1, N-1,\\
-E_iI_i\hat{w}''(x_i^-) + E_{i+1}I_{i+1}\hat{w}''(x_i^+), & i = 1, N-1,\\
E_NI_N\hat{w}'''(L),\\
-E_NI_N\hat{w}''(L).
\end{cases} \tag{2.6}
$$

LEMMA 2.1. *$\mathcal{F}_{\mu w}(0, 0, 0): \mathbb{C} \times \mathcal{W} \to H \times \mathbb{C}^{2N}$ is an isomorphism.*

Proof. It is enough to establish that for all $f \in H$, and $(\alpha_i, \beta_i) \in \mathbb{C}^2$, $i = 1, N$, the following system

$$
\mathcal{F}_{\mu w}(0,0,0)\hat{\mu}\hat{w} = \begin{pmatrix} f \\ \alpha_i \\ \beta_i \end{pmatrix} \text{ admits a unique solution } \hat{\mu}, \hat{w} \text{ that depends continuously on the data } f, \alpha_i, \beta_i. \tag{2.7}
$$

Consider (2.7), with (2.6) in mind. Let $g \in W$ be the unique solution of

$$
\begin{aligned}
&g'''' = 0, \quad x_{i-1} < x < x_i, \quad i = 1, N, \\
&E_i I_i g'''(x_i^-) - E_{i+1} I_{i+1} g'''(x_i^+) = \alpha_i, \qquad i = 1, N-1, \\
&-E_i I_i g''(x_i^-) + E_{i+1} I_{i+1} g''(x_i^+) = \beta_i, \qquad i = 1, N-1, \\
&E_N I_N g'''(L) = \alpha_N, \\
&-E_N I_N g''(L) = \beta_N.
\end{aligned} \tag{2.8}
$$

g is the minimizer of the strain energy plus the potential energy of the forces α_i and moments β_i:

$$
\frac{1}{2} \sum_{i=1}^{N} \left[E_i I_i \int_{x_{i-1}}^{x_i} v''(x)^2 \, dx + \alpha_i v(x_i) + \beta_i v'(x_i) \right],
$$

g is piecewise cubic and depends continuously on the data α_i and β_i.

We set $\hat{w} = g + \hat{v}$.

Then $\hat{v}$ satisfies

$$
\begin{aligned}
&E_i I_i \hat{v}'''' + \nu_p^2 m_i (g + \hat{v}) + 2\nu_p m_i \hat{\mu} \phi_p = f, \quad x_{i-1} < x < x_i, \quad i = 1, N, \\
&E_i I_i \hat{v}'''(x_i^-) - E_{i+1} I_{i+1} \hat{v}'''(x_i^+) = 0, \qquad i = 1, N-1, \\
&-E_i I_i \hat{v}''(x_i^-) + E_{i+1} I_{i+1} \hat{v}''(x_i^+) = 0, \qquad i = 1, N-1, \\
&E_N I_N \hat{v}'''(L) = 0, \\
&-E_N I_N \hat{v}''(L) = 0;
\end{aligned} \tag{2.9}
$$

equivalently,

$$
A\hat{v} + \nu_p^2 M \hat{v} = f - \nu_p^2 M g - 2\nu_p \hat{\mu} M \phi_p \tag{2.9}
$$

where A is an unbounded self-adjoint operator on H, with A^{-1} compact. By the Fredholm alternative, (2.9) admits a solution if and only if the right-hand side of (2.9) is orthogonal to ϕ_p; this gives a unique $\hat{\mu}$

$$
\hat{\mu} = \frac{(f - \nu_p^2 M g, \phi_p)}{2\nu_p (M\phi_p, \phi_p)}, \tag{2.10}
$$

which is continuous with respect to $f \in H$, $g \in H$.

Denote by $\mathcal{T}$ the pseudoinverse of $A + \nu_p^2 M$; $\mathcal{T}$ is continuous from $\{\phi_p\}^\perp$ onto $\mathcal{W}$. Then the general solution of (2.9) is given by

$$
\hat{v} = \mathcal{T} h + \lambda \phi_p \quad \text{where } h = f - \nu_p^2 M g - 2\nu_p \hat{\mu} M \phi_p, \quad \lambda \in \mathbb{C}.
$$

Since $\hat{w} = \hat{v} + g \in \mathcal{W}$, we have $(\mathcal{T} h + \lambda \phi_p + g, \phi_p) = 0$, which gives a unique $\lambda = -(g, \phi_p)/(\phi_p, \phi_p)$, hence globally, a unique $\hat{w}$ that depends continuously on the data. □

THEOREM 2.2. *For any fixed $p \in \mathbb{N}$, (2.2) has, for K small enough, solutions of the form*

$$
\begin{aligned}
&\nu_{\pm p}(K) = \pm i\omega_p + \hat{\mu}_p(K) + |K| \varepsilon_{\pm p}(K), \\
&\phi_{\pm p}(K) = \phi_p \pm \hat{w}_p(K) + |K| \eta_{\pm p}(K)
\end{aligned}
$$

with

$$
\hat{\mu}_p(K) = -\frac{1}{2} \frac{(KC\phi_p, C\phi_p)}{(M\phi_p, \phi_p)}, \qquad \hat{w}_p(K) \in \mathcal{W}
$$

linear continuous in K and $\varepsilon_{\pm p}(K) \in \mathbb{C}$, $\eta_{\pm p}(K) \in \mathcal{W}$ *regular functions defined for K small enough.*

Proof. The existence of the functions follows from the Implicit Function Theorem applied to $\mathcal{F}$ near $(\pm i\omega_p, \phi_p)$ and valid by Lemma 2.1. We note that $\hat{\mu} = \hat{\mu}_{\pm p}(K)$, $\hat{w} = \hat{w}_{\pm p}(K)$ are solutions of

$$\mathcal{F}_{\hat{\mu}\hat{w}}(0,0,0)\hat{\mu}\hat{w} = -\mathcal{F}_K(0,0,0)K = \begin{cases} 0 \\ \nu_p K_i \begin{pmatrix} \phi_p(x_i) \\ \phi_p'(x_i) \end{pmatrix}, \end{cases} \qquad i = 1, N, \tag{2.11}$$

which is just a special form of (2.7) with $f = 0$. Hence

$$\hat{\mu} = -\frac{\nu_p}{2} \frac{(Mg, \phi_p)}{(M\phi_p, \phi_p)}. \tag{2.12}$$

By (2.11) and (2.12), changing ν_p into $-\nu_p$ gives the opposite sign for g, hence the same $\hat{\mu}$.

On the other hand, since $\hat{w} = \mathcal{T}h + \lambda\phi_p + g$ with $h = -\nu_p^2 Mg - 2\nu_p \hat{\mu} M\phi_p$, the signs of g, h and $\lambda = -(g, \phi_p)/(\phi_p, \phi_p)$ change and so $\hat{w}$ changes sign.

So the only thing not yet established is the expression of $\hat{\mu}_p(K)$ given by the theorem.

We start from (2.12) $(Mg, \phi_p) = (g, M\phi_p) = -(1/\nu_p^2)(g, A\phi_p)$ by (2.3) $= -(1/\nu_p^2)(Ag, \phi_p)$. We note that (2.8) can be written as $Ag + C^*\binom{\alpha_i}{\beta_i} = 0$, thus

$$(Mg, \phi_p) = \frac{1}{\nu_p^2}\left(C^*\begin{pmatrix}\alpha_i \\ \beta_i\end{pmatrix}, \phi_p\right) = \frac{1}{\nu_p^2}\left(\begin{pmatrix}\alpha_i \\ \beta_i\end{pmatrix}, C\phi_p\right)_{\mathbb{R}^{2N}}$$

with

$$\begin{pmatrix}\alpha_i \\ \beta_i\end{pmatrix} = \nu_p K_i \begin{pmatrix}\phi_p(x_i) \\ \phi_p'(x_i)\end{pmatrix},$$

thus $(Mg, \phi_p) = (1/\nu_p)(KC\phi_p, C\phi_p)_{\mathbb{R}^{2N}}$ and the result follows from (2.12). □

Remark. The method also works for a beam with variable mass density and flexural rigidity, clamped at the left end and controlled at the right end. The result is exactly the same.

3. Application to a single beam. We consider a single beam with length L, constant mass density m, and flexural rigidity EI:

$$\begin{aligned} &m\ddot{y} + EIy'''' = 0, \qquad 0 < x < L, \\ &y(0,t) = y'(0,t) = 0, \\ &EI\begin{pmatrix} y'''(L,t) \\ -y''(L,t)\end{pmatrix} = K\begin{pmatrix}\dot{y}(L,t) \\ \dot{y}'(L,T)\end{pmatrix}. \end{aligned} \tag{3.1}$$

It is known [13] that for the uncontrolled system, $\nu_p = \pm i\omega_p$ with

$$\omega_p = \sqrt{EI/mL^4}\,\alpha_p^2 \quad \text{where the } \alpha_p,\ p = 1, 2, \cdots, \text{ are the positive solutions of} \tag{3.2}$$

$$1 + \cos\alpha \operatorname{ch}\alpha = 0. \tag{3.3}$$

Moreover,

$$\alpha_p = (p - \tfrac{1}{2})\pi + \delta_p \quad \text{where} \quad |\delta_p| \leqq 2C_p e^{-(p-1/2)\pi}, \quad C_p < 2 \text{ for } p > 1 \text{ [1]}. \tag{3.4}$$

The $L^2(0, L)$-normalized eigenfunctions are

$$\phi_p(x) = \frac{1}{\sqrt{L}}[\operatorname{ch} z - \cos z - \gamma(\alpha)(\operatorname{sh} z - \sin z)] \tag{3.5}$$

where $\alpha = \alpha_p$, $z = \alpha x/L$ and $\gamma(\alpha) = (\operatorname{ch} \alpha + \cos \alpha)/(\operatorname{sh} \alpha + \sin \alpha)$.

From (3.3) we easily get

$$\phi_p(L) = \frac{2(-1)^{p+1}}{\sqrt{L}}$$

and also

$$\phi_p'(L) = \frac{2\alpha}{L^{3/2}}\left(\operatorname{tg}\frac{\alpha}{2}\right)^{(-1)^p}.$$

With the form of matrix K considered in [2] $K = \begin{pmatrix} k_0 & -\gamma \\ \gamma & k_1 \end{pmatrix}$, the first-order approximations of the eigenvalues of the generator of (3.1) given by Theorem 2.2 are

$$\nu_p(K) = \pm i\sqrt{\frac{EI}{mL^4}}\,\alpha_p^2 - \frac{2}{m}\left[\frac{k_0}{L} + \frac{k_1}{L^3}\,\alpha_p^2\left(\operatorname{tg}^2\frac{\alpha_p}{2}\right)^{(-1)^p}\right] + \cdots. \tag{3.6}$$

Now, if we use (3.4) noting that δ_p is very small as soon as $p \geqq 2$, we have the sharp approximation

$$\nu_p(K) = \pm i\sqrt{\frac{EI}{mL^4}}\,\alpha_p^2 - \frac{2}{m}\left[\frac{k_0}{L} + \frac{k_1}{L^3}\,\alpha_p^2\right] + \cdots,$$

which gives the first-order evolution of the spectrum with respect to K, as shown in Fig. 2.

This is completely consistent with the conclusion obtained in [2] and [10] from a numerical experiment, at least for the first eigenvalues.

Remark. Higher-order expansions are available using a system of formal calculus such as REDUCE to solve the characteristic equation for the feedback system.

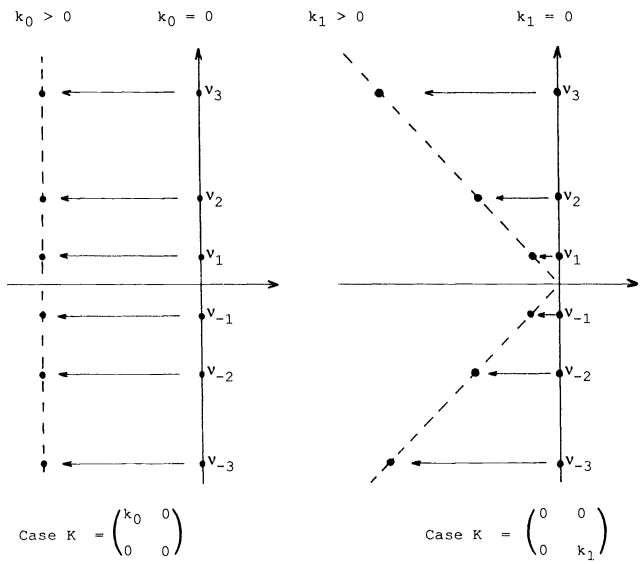


FIG. 2. *Sensitivity of the spectrum with respect to feedback coefficients.*

4. Obtaining an optimal rate of decay. The expansion of the eigenvalue $\nu_p(K)$ with respect to K given in Theorem 2.2 is valid for any fixed $p \in \mathbb{Z}$. In general, the higher-order term in the expansion is not controlled uniformly in p. Consider, for instance, the case of the single beam of § 3, with the matrix

$$K = \begin{pmatrix} k_0 & -\gamma \\ \gamma & k_1 \end{pmatrix}, \qquad k_1 \neq 0.$$

It has been proved that the eigenvalues of large modulus of that system admit a vertical asymptote [3]-[5], [10]. Thus an expansion of the form (3.6), $\nu_p(K) = \pm i(EI/mL^4)^{1/2}\alpha_p^2 - 2/m[k_0/L + k_1\alpha_p^2/L^3] + |K|\varepsilon_p(K)$ with $\varepsilon_p(K) \to 0$ as $K \to 0$ *uniformly in p*, cannot hold if $k_1 \neq 0$.

However, following an idea used in [12], we will get the uniformity, and in fact, very sharp estimates of the higher-order terms, in the particular case of a single beam with control on the shear only ($k_1 = \gamma = 0$). Thus we consider in this section the system

$$\begin{aligned} &m\ddot{y} + EIy'''' = 0, \qquad 0 < x < L, \\ &y(0, t) = y'(0, t) = 0, \\ &EIy'''(L, t) = k_0\dot{y}(L, t), \\ &-EIy''(L, t) = 0. \end{aligned} \tag{4.1}$$

We normalize L and set $a^4 = EI/mL^4$, $k = k_0L^3/EI$; then (4.1) rewrites as

$$\begin{aligned} &\ddot{y} + a^4y'''' = 0, \qquad 0 < x < 1, \\ &y(0, t) = y'(0, t) = 0, \\ &y'''(1, t) = k\dot{y}(1, t), \\ &y''(1, t) = 0. \end{aligned} \tag{4.2}$$

We will even suppose $a = 1$ with a new time scale (keeping the same notation for the new k).

The notation is the same as in §§ 2 and 3: $H = L^2(0, 1)$, $V = \{v \in H^2(0, 1) \mid v(0) = v'(0) = 0\}$, $\langle Au, v\rangle = \int_0^1 u''(x)v''(x)\,dx$, $A \in \mathcal{L}(V, V')$, and, as an unbounded operator on H, $\operatorname{dom} A = W = \{w \in H^4(0, 1) \cap V \mid w''(1) = w'''(1) = 0\}$; here $K = \left(\begin{smallmatrix} k & 0 \\ 0 & 0 \end{smallmatrix}\right)$.

Although not necessary, it is more convenient to work on $H \oplus H$ instead of $V \oplus H$. Since the form $\langle Au, v\rangle$ is V-elliptic, A admits a square root [8], which is a self-adjoint positive unbounded operator $A^{1/2}$ with domain V.

We set $z_1 = A^{1/2}y$, $z_2 = \dot{y}$ and (1.3), and hence (4.2) rewrites as

$$\begin{pmatrix} \dot{z}_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 0 & A^{1/2} \\ -A^{1/2} & -C^*KC \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \mathcal{A}(K) \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \tag{4.3}$$

now with a generator $\mathcal{A}(K)$ (same notation as before, but not same operator) on $H \oplus H$.

For the moment we follow the methodology of § 2. Let $\nu \in \mathbb{C}$ be an eigenvalue of $\mathcal{A}(K)$, $\left(\begin{smallmatrix} \phi \\ \psi \end{smallmatrix}\right) \in \operatorname{dom} \mathcal{A}(K)$ be an eigenvector of $\mathcal{A}(K)$, i.e.,

$$A^{1/2}\psi = \nu\phi, \tag{4.4}$$

$$-A^{1/2}\phi - C^*KC\psi = \nu\psi, \quad \text{or else} \tag{4.5}$$

$$A\psi + \nu C^*KC\psi + \nu^2\psi = 0. \tag{4.6}$$

So we must solve (4.6), the same equation as in § 2, and then $\begin{pmatrix} A^{1/2}\psi \\ \nu\psi \end{pmatrix}$ is an eigenvector of $\mathcal{A}(K)$.

4.1. Uncontrolled system ($K=0$).

$$A\psi + \nu^2\psi = 0. \tag{4.7}$$

According to § 3, we get the sequence of eigenvalues $(\nu_p)_{p\in\mathbb{Z}^*}$ with $\nu_{\pm p} = \pm i\alpha_p^2$, $p>0$ where the α_p are solutions of (3.3) and the sequence of eigenvectors $\psi_{\pm p}(x) = \phi_p(x)$, $p>0$ given by (3.5).

Thus ν_p, $p\in\mathbb{Z}^*$ is the set of eigenvalues of $\mathcal{A}(0)$ and, by (4.4) $\xi_p = 1/\sqrt{2}\begin{pmatrix} \psi_p \\ \varepsilon i\psi_p \end{pmatrix}$, $p\in\mathbb{Z}^*$ ($\varepsilon=\pm 1$ whether $p>0$ or $p<0$) are normalized associated eigenvectors. Moreover, the ξ_p constitute a complete orthonormal basis of $H\oplus H$ (in the sequel, all the spaces are implicitly complexified).

4.2. Feedback system ($K\neq 0$). Now p is fixed, say >0 for convenience in the subsequent computations.

By the results of § 2, we immediately have the following lemma.

LEMMA 4.1. *For any fixed p, and k small enough,* (4.6) *admits solutions of the form*

$$\nu_p(k) = \nu_p - 2k + k\varepsilon_p(k), \qquad p\in\mathbb{Z}^* \tag{4.8}$$

$$\psi_p(k) = \psi_p + kz_p + k\eta_p(k), \qquad p\in\mathbb{Z}^* \tag{4.9}$$

where $\lim_{k\to 0}\varepsilon_p(k) = 0$ *in* $\mathbb{C}$, $z_p\in\mathcal{W}$, *and* $\lim_{k\to 0}\eta_p(k) = 0$ *in* $\mathcal{W}$.

$\mathcal{W}$ is the orthogonal of ψ_p in W for the H-scalar product.

We recall that $\psi_p = \phi_p$ as defined in § 3.

LEMMA 4.2. *$z_p\in\mathcal{W}$ is given by*

$$z_p(x) = \nu_p\psi_p(1)\sum_{q\neq p}\frac{\psi_q(1)}{\nu_q^2-\nu_p^2}\psi_q(x) \tag{4.10}$$

where the series converges uniformly in x. Moreover, $\|z_p\|_{c^0}\leqq M|\log p|/|p|$.

Proof. According to § 2, kz_p is the unique solution of (2.11) that we rewrite here as ($\hat{\mu} = -2k$)

$$\begin{aligned} &z_p'''' + \nu_p^2 z_p - 4\nu_p\psi_p = 0, \qquad z_p\in\mathcal{W}, \\ &z_p'''(1) = \nu_p\psi_p(1), \\ &z_p''(1) = 0. \end{aligned} \tag{4.11}$$

Proceeding as in the proof of Lemma 2.1, we get

$$z_p = g - \frac{(g,\psi_p)}{(\psi_p,\psi_p)}\psi_p + \mathcal{T}(-\nu_p^2 g + 4\nu_p\psi_p)$$

where g is the solution of

$$\begin{aligned} &g'''' = 0, \\ &g(0) = g'(0) = 0, \\ &g'''(1) = \nu_p\psi_p(1), \\ &g''(1) = 0; \end{aligned}$$

therefore $g(x) = \nu_p\psi_p(1)[x^3/6 - x^2/2]$.

$\mathcal{T}$ is the pseudoinverse of $A+\nu_p^2$ defined from $\{\psi_p\}^\perp$ onto $\mathcal{W}$, thus for all $h\in H$, $h\perp\psi_p$:

$$\mathcal{T}h=\sum_{\substack{q=1\\q\neq p}}^{\infty}\frac{(h,\psi_q)}{\nu_p^2-\nu_q^2}\psi_q,$$

so that

$$\begin{aligned}z_p&=\sum_{\substack{q\neq p\\q>0}}(g,\psi_q)\psi_q+\sum_{\substack{q\neq p\\q>0}}\frac{(-\nu_p^2 g+4\nu_p\psi_p,\psi_q)}{\nu_p^2-\nu_q^2}\psi_q\\&=-\sum_{\substack{q\neq p\\q>0}}\frac{\nu_q^2(g,\psi_q)}{\nu_p^2-\nu_q^2}\psi_q=\sum_{\substack{q\neq p\\q>0}}\frac{(g,\psi_q'''')}{\nu_p^2-\nu_q^2}\psi_q\\&=-\sum_{q\neq p}\frac{g'''(1)\psi_q(1)}{\nu_p^2-\nu_q^2}\psi_q\end{aligned}$$

after integration by parts, thus $z_p=\nu_p\psi_p(1)\sum_{q\neq p}\psi_q(1)/(\nu_q^2-\nu_p^2)\psi_q$ with convergence in the sense of $L^2(0,1)$ and (4.10) is established. Since the ψ_q are bounded in L^∞-norm uniformly in p (Lemma 1, Appendix) the convergence is also uniform. From the boundedness of the ψ_q and the estimates on the series $1/(\nu_q^2-\nu_q^2)$ (Lemma 2, Appendix) the estimate on $\|z_p\|_{c^0}$ follows. □

Now we want to get sharp estimates on the higher-order terms in (4.8) and (4.9). So we inject the expressions of $\nu_p(k)$ and $\psi_p(k)$ into (4.6), replacing $\varepsilon_p(k)$ and $\eta_p(k)$ by ε and η for simplicity; after the simplifications due to the first-order approximations, we get

$$\begin{aligned}&A\eta+\nu_p^2\eta+(4k+k\varepsilon^2+2\nu_p\varepsilon-4k\varepsilon)(\psi_p+kz_p+k\eta)\\&\quad-4\nu_pk(z_p+\eta)+(\varepsilon-2)C^*KC(\psi_p+kz_p+k\eta)+\nu_pC^*KC(z_p+\eta)=0\end{aligned}\tag{4.12}$$

or, in strong form:

$$\begin{aligned}&\eta''''+\nu_p^2\eta=4\nu_pk(z_p+\eta)-(4k+k\varepsilon^2+2\nu_p\varepsilon-4k\varepsilon)(\psi_p+kz_p+k\eta),\\&\eta'''(1)=(\varepsilon-2)k(\psi_p(1)+kz_p(1)+k\eta(1))+\nu_pk(z_p(1)+\eta(1)),\\&\eta''(1)=0.\end{aligned}\tag{4.13}$$

We denote the right-hand side of (4.13) by $(f,\alpha,0)$ and apply the same technique as in Lemma 2.1 of § 2.

Let $\eta=\hat\eta+g$ where g absorbs the nonzero boundary condition

$$g(x)=\frac{\alpha}{6}x^3-\frac{\alpha}{2}x^2.$$

Then (4.13) rewrites as

$$\begin{aligned}&\hat\eta''''+\nu_p^2\hat\eta=f-\nu_p^2g,\\&\hat\eta'''(1)=0,\\&\hat\eta''(1)=0,\end{aligned}\tag{4.14}$$

and (4.14) has a solution if and only if $(f-\nu_p^2g,\psi_p)=0$, that is,

$$\nu_p^2(g,\psi_p)=(f,\psi_p)=-4k+k\varepsilon^2+2\nu_p\varepsilon-4k\varepsilon.\tag{4.15}$$

We have already proved in the course of Lemma 4.2 that

$$(g, \psi_p) = \frac{1}{\nu_p^2} g'''(1)\psi_p(1) = \frac{1}{\nu_p^2} \alpha \psi_p(1)$$

so that (4.15) is equivalent to

$$\begin{aligned} &-4k + k\varepsilon^2 + 2\nu_p\varepsilon - 4k\varepsilon \\ &\quad = \psi_p(1)[(\varepsilon - 2)k(\psi_p(1) + kz_p(1) + k\eta(1)) + \nu_p k(z_p(1) + \eta(1))], \end{aligned} \tag{4.16}$$

i.e.,

$$\varepsilon = \frac{k}{2\nu_p}[4\varepsilon + 4 - \varepsilon^2 + \psi_p(1)(\varepsilon - 2)(\psi_p(1) + kz_p(1) + k\eta(1))] + \frac{k}{2}[z_p(1) + \eta(1)]. \tag{4.17}$$

Once (4.17) is satisfied, the unique solution of (4.13) is given by (see again the proof of Lemma 4.2)

$$\eta = g - (g, \psi_p)\psi_p + \mathcal{T}(f - \nu_p^2 g) = \sum_{\substack{q>0 \\ q\neq p}} (g, \psi_q)\psi_q + \sum_{\substack{q>0 \\ q\neq p}} \frac{(f - \nu_p^2 g, \psi_q)}{\nu_p^2 - \nu_q^2} \psi_q$$

$$= \sum_{\substack{q>0 \\ q\neq p}} \frac{-\nu_q^2 (g, \psi_q)}{\nu_p^2 - \nu_q^2} \psi_q + \sum_{\substack{q>0 \\ q\neq p}} \frac{(f, \psi_q)}{\nu_p^2 - \nu_q^2} \psi_q = \sum_{\substack{q>0 \\ q\neq p}} \frac{g'''(1)\psi_q(1)}{\nu_q^2 - \nu_p^2} \psi_q + \sum_{\substack{q>0 \\ q\neq p}} \frac{(f, \psi_q)}{\nu_p^2 - \nu_q^2} \psi_q,$$

$$\begin{aligned} \eta = k \sum_{\substack{q>0 \\ q\neq p}} &\frac{(\varepsilon - 2)(\psi_p(1) + kz_p(1) + k\eta(1)) + \nu_p(z_p(1) + \eta(1))}{\nu_q^2 - \nu_p^2} \psi_q(1)\psi_q \\ &+ k \sum_{\substack{q>0 \\ q\neq p}} \frac{4\nu_p - (4k + k\varepsilon^2 + 2\nu_p\varepsilon - 4k\varepsilon)}{\nu_p^2 - \nu_q^2} (z_p + \eta, \psi_q)\psi_q. \end{aligned} \tag{4.18}$$

Finally, (ε, η) satisfies (4.13) if and only if

$$\varepsilon = kF_p(k, \varepsilon, \eta), \tag{4.19}$$

$$\eta = kG_p(k, \varepsilon, \eta) \tag{4.20}$$

where kF_p and kG_p are the right-hand sides of (4.17) and (4.18), respectively.

THEOREM 4.3. *The functions ε_p and η_p of Lemma 4.1 can be defined on an interval $(-\alpha, \alpha)$ in k independently of p; moreover,*

$$|\varepsilon_p(k)| = \frac{|\log p|}{|p|} O(k), \tag{4.21}$$

$$\|\eta_p(k)\|_{C^0} = \frac{|\log p|^2}{|p|^2} O(k) \tag{4.22}$$

where $O(k)$ is uniform in p.

Proof. We consider (4.19), (4.20) as a fixed-point formulation with parameters k and p

$$(\varepsilon, \eta) = T_p(k, \varepsilon, \eta). \tag{4.23}$$

Let $B = \{k, \varepsilon, \eta: |k| \leqq 1, |\varepsilon| \leqq 1, \|\eta\|_{C^0} \leqq 1\} \subset \mathbb{R} \times \mathbb{C} \times C^0([0, 1])$.

From the expressions of F_p and G_p, Lemma 4.2, and Lemmas 1-3 of the Appendix, we have (M denotes generic constants) on B:

$$\begin{aligned} |F_p(k,\varepsilon,\eta)| &\leqq \frac{M}{|\nu_p|}+M\frac{|\log p|}{|p|}+M\|\eta\|_{c^0} \\ &\leqq M\frac{|\log p|}{|p|}+M\|\eta\|_{c^0}, \end{aligned} \tag{4.24}$$

$$\begin{aligned} \|G_p(k,\varepsilon,\eta)\|_{c^0} &\leqq M\frac{|\log p|}{|p|^3}+M\frac{|\log p|^2}{|p|^2}+M\frac{|\log p|}{|p|}\|\eta\|_{c^0} \\ &\leqq M\frac{|\log p|^2}{|p|^2}+M\frac{|\log p|}{|p|}\|\eta\|_{c^0}, \end{aligned} \tag{4.25}$$

all the constants being independent of p.

Similarly, we get, for the partial derivatives with respect to ε and η:

$$\left|\frac{\partial F_p}{\partial\varepsilon}(k,\varepsilon,\eta)\right| \leqq \frac{M}{|\nu_p|} \leqq M\frac{|\log p|}{|p|}, \tag{4.26}$$

$$\left|\frac{\partial F_p}{\partial\eta}(k,\varepsilon,\eta)\right| \leqq M,$$

$$\left\|\frac{\partial G_p}{\partial\varepsilon}(k,\varepsilon,\eta)\right\|_{c^0} \leqq M\frac{|\log p|}{|p|}, \tag{4.27}$$

$$\left\|\frac{\partial G_p}{\partial\eta}(k,\varepsilon,\eta)\right\|_{c^0} \leqq M\frac{|\log p|}{|p|}.$$

From (4.26) and (4.27) it follows that, for $|k| \leqq \alpha$ small enough, T is a contraction, uniformly in p (and k, for $|k|$ small).

By classical fixed point theory, we get the functions ε and η defined on an interval $(-\alpha, \alpha)$, uniformly in p.

From (4.25), we get $\|\eta_p(k)\|_{c^0} \leqq Mk(|\log p|/|p|)$ and, by bootstrap, we get (4.22).

Finally, (4.21) follows from (4.24) and (4.22). □

Remark. For a general feedback matrix with $k_1 \neq 0$, we must add terms involving $\psi_p'(1)$, which is $O(p)$. Roughly speaking, we lose at least a factor $1/p$ in the series so that estimates (4.21) and (4.22) for ε_p and η_p cannot be better than $\log|p|$ using our technique. This is consistent with the remark concerning uniformity with respect to p at the beginning of this section.

THEOREM 4.4. *For k small enough, $\mathcal{A}(K)$ admits a sequence of eigenvectors that constitute a Riesz basis in $H \oplus H$.*

Proof. We consider the eigenvalues $\nu_p(k)$ and eigenvectors $\psi_p(k)$ given by Lemma 4.1 and Theorem 4.3.

According to (4.4), (4.6), $\binom{A^{1/2}\psi_p(k)}{\nu_p(k)\psi_p(k)}$ is an eigenvector of $\mathcal{A}(K)$. We normalize it by considering

$$\begin{aligned} \xi_p(k) &= \frac{1}{\sqrt{2}\,\omega_p}\begin{pmatrix} A^{1/2}\psi_p + kA^{1/2}(z_p+\eta_p) \\ (\nu_p - 2k + k\varepsilon_p)(\psi_p + k(z_p+\eta_p)) \end{pmatrix} \\ &= \xi_p + \frac{k}{\sqrt{2}\,\omega_p}\begin{pmatrix} A^{1/2}(z_p+\eta_p) \\ (\varepsilon_p - 2)\psi_p + (\nu_p - 2k + k\varepsilon_p)(z_p+\eta_p) \end{pmatrix} \\ &= \xi_p + \frac{k}{\sqrt{2}\,\omega_p}\,h_p. \end{aligned}$$

By Lemma 4.2, $\|z_p\|_{c^0} \leqq M(|\log p|/|p|)$ and by (4.10), $A^{1/2}z_p = \nu_p\psi_p(1)\sum_{q\neq p,q>0}(\psi_q(1)/(\nu_q^2-\nu_p^2))\omega_q\psi_q(x)$, thus by Lemma 3 of the Appendix,

$$\|A^{1/2}z_p\|_{c^0} \leqq M|\nu_p|\frac{|\log p|}{|p|} = M\omega_p\frac{|\log p|}{|p|}.$$

Similarly, using (4.18) and expanding $A^{1/2}\eta$, we obtain the same estimates on $\|\eta_p\|_{c^0}$ and $\|A^{1/2}\eta_p\|_{c^0}$ up to k, which is bounded.

Finally, we obtain that

$$\left\|\frac{k}{\sqrt{2}\,\omega_p}h_p\right\|_{H\oplus H} \leqq \frac{k}{\sqrt{2}}\left\|\frac{h_p}{\omega_p}\right\|_{c^0\times c^0} \leqq Mk\left|\frac{\log p}{p}\right|$$

so that

$$\sum_{p\in\mathbb{Z}^*}|\xi_p(k)-\xi_p|^2_{H\oplus H} \leqq Mk^2\sum_{\mathbb{Z}^*}\left|\frac{\log p}{p}\right|^2 \leqq Mk^2 < 1$$

for k small enough.

According to a theorem of Paley and Wiener [11, p. 206] the $\xi_p(k)$ are a Riesz basis of $H\oplus H$. □

Remark 1. This result also proves that by our perturbation method, we have obtained all the eigenvalues of $\mathcal{A}(K)$.

Remark 2. Alternatively, we have proved that the eigenvectors

$$\frac{1}{\sqrt{2}\,\omega_p}\begin{pmatrix}\psi_p(k)\\ \nu_p\psi_p(k)\end{pmatrix}$$

form a Riesz basis in $V\oplus H$.

COROLLARY 4.5. *For k small enough, the spectral radius of the semigroup* $S_K(t)$ *is given by*

$$\omega(k) = \sup_{p\in\mathbb{Z}^*}\{\operatorname{Re}\nu_p(k)\} = -2k+O(k).$$

Proof. The first equality is a direct consequence of Theorem 4.4 (or Remark 2), the second follows from the uniformity of the estimates of ε_p. □

Remark. A general study of the asymptotic distribution of the eigenfrequencies, either for a single beam with active control at one end, or for two- or N-coupled beams with various dissipative joint conditions, has been carried out in [3]-[5]. Those results and the sensitivity analysis given here in § 2 are mutually complementary. Our results are essentially valid even at low frequencies.

Moreover, the asymptotic gaps for the eigenvalues given in [4] or [5] are one of the general assumptions that make the method of § 4 work, together with boundedness properties of the eigenfunctions.

Thus it seems possible to extend the results of § 4 to systems of coupled beams, provided suitable information on the eigenfunctions is available.

Appendix. Here we prove the technical estimates that are used in the proofs of Lemma 4.2, and Theorems 4.3 and 4.4.

The notation is that of §§ 3 and 4. All the constants m, EI, L are equal to 1, without loss of generality; M denotes constants.

LEMMA 1. $\|\phi_p\|_{c^0} \leqq M$; $\|\phi_p'\|_{c^0} \leqq M\alpha_p$ *where* $\alpha_p > 0$ *is a solution of* (3.3).

Proof. We recall that $\phi_p(x) = \operatorname{ch}\alpha x - \cos\alpha x - \gamma(\alpha)[\operatorname{sh}\alpha x - \sin\alpha x]$

$$\gamma(\alpha) = \frac{-\sin\alpha}{(-1)^p+\cos\alpha}, \qquad \alpha = \alpha_p.$$

By [1] $\alpha=(p-\frac{1}{2})\pi+2c_p(-1)^{p-1}\,e^{-(p-1/2)\pi}=\tilde{\alpha}+\varepsilon$ and $c_p<2$ when $p>1$. Thus for large p

$$\sin\alpha=(-1)^{p-1}+\eta,\qquad |\eta|\leqq\frac{\varepsilon^2}{2},$$

$$|\cos\alpha|\leqq|\varepsilon|\Rightarrow\gamma(\alpha)=1+\xi\quad\text{with } |\xi|\leqq 2\varepsilon\leqq 4c_p\,e^{-\tilde{\alpha}}.$$

Therefore $\phi_p(x)=\operatorname{ch}\alpha x-(1+\xi)\operatorname{sh}\alpha x$

$$-\cos\alpha x+(1+\xi)\sin\alpha x\Rightarrow|\phi_p(x)|\leqq-\xi\frac{e^{\alpha x}}{2}+\left(1+\frac{\xi}{2}\right)e^{-\alpha x}+M\leqq M.$$

For the estimate of the derivative

$$\frac{\phi_p'(x)}{\alpha}=\operatorname{sh}\alpha x+\sin\alpha x-\gamma(\alpha)[\operatorname{ch}\alpha x-\cos\alpha x]$$

$$=\operatorname{sh}\alpha x-(1+\xi)\operatorname{ch}\alpha x+\sin\alpha x+(1-\xi)\cos\alpha x.$$

Then, proceeding as before, we get $|\phi_p'(x)/\alpha|\leqq M$. □

Remark. According to the expressions of $\phi_p(1)$ and $\phi_p'(1)$ given in § 3, the estimates of Lemma 1 are optimal.

LEMMA 2.

$$\sum_{\substack{q>0\\ q\neq p}}\frac{1}{|\nu_p^2-\nu_q^2|}\leqq M\frac{|\log p|}{|p|^3}.$$

Proof.

$$\sum_{\substack{q>0\\ q\neq p}}\frac{1}{|\nu_p^2-\nu_q^2|}=\sum_{\substack{q>0\\ q\neq p}}\frac{1}{|\alpha_q^4-\alpha_p^4|}$$

$$\leqq\frac{1}{\alpha_p^4-\alpha_{p-1}^4}+\frac{1}{\alpha_{p+1}^4-\alpha_p^4}+\int_{\alpha_1}^{\alpha_{p-1}}\frac{dx}{\alpha_p^4-x^4}+\int_{\alpha_{p+1}}^{\infty}\frac{dx}{x^4-\alpha_p^4}$$

$$\leqq\frac{M}{p^3}+\frac{1}{\alpha_p^2}\left[\int_{\alpha_1}^{\alpha_{p-1}}\frac{dx}{\alpha_p^2-x^2}+\int_{\alpha_{p+1}}^{\infty}\frac{dx}{x^2-\alpha_p^2}\right]$$

$$\leqq\frac{M}{p^3}+\frac{M}{p^3}\left(\left[\log\frac{\alpha_p+x}{\alpha_p-x}\right]_{\alpha_1}^{\alpha_{p-1}}-\left[\log\frac{x+\alpha_p}{x-\alpha_p}\right]_{\alpha_{p+1}}^{\infty}\right)$$

$$\leqq\frac{M}{p^3}+\frac{M}{p^3}\log\frac{(\alpha_p+\alpha_{p-1})(\alpha_p-\alpha_1)(\alpha_p+\alpha_{p+1})}{(\alpha_p-\alpha_{p-1})(\alpha_p+\alpha_1)(\alpha_{p+1}-\alpha_p)}\leqq M\frac{|\log p|}{|p|^3}.$$ □

Remark. Using also the comparison of series and integrals, a similar reverse inequality can be established. Therefore, the estimate is optimal.

LEMMA 3.

$$\sum_{\substack{q>0\\ q\neq p}}\frac{|\nu_q|}{|\nu_p^2-\nu_q^2|}\leqq M\frac{|\log p|}{|p|}.$$

Proof.

$$\sum_{\substack{q>0\\ q\neq p}} \frac{|\nu_q|}{|\nu_p^2-\nu_q^2|} = \sum_{\substack{q>0\\ q\neq p}} \frac{\alpha_q^2}{|\alpha_q^4-\alpha_p^4|}$$

$$\leqq \frac{\alpha_{p-1}^2}{\alpha_p^4-\alpha_{p-1}^4} + \frac{\alpha_{p+1}^2}{\alpha_{p+1}^4-\alpha_p^4} + \int_{\alpha_1}^{\alpha_{p-1}} \frac{x^2\,dx}{\alpha_p^4-x^4} + \int_{\alpha_{p+1}}^{\infty} \frac{x^2}{x^4-\alpha_p^4}\,dx$$

$$\leqq \frac{M}{p} + \frac{1}{2}\int_{\alpha_1}^{\alpha_{p-1}} + \int_{\alpha_{p+1}}^{\infty} \left|\frac{1}{\alpha_p^2-x^2} - \frac{1}{\alpha_p^2+x^2}\right| dx$$

$$\leqq \frac{M}{p} + \frac{1}{2\alpha_p}\left[\frac{1}{2}\log\left|\frac{\alpha_p+x}{\alpha_p-x}\right| + \operatorname{arc\,tg}\frac{x}{\alpha_p}\right]_{\alpha_1}^{\alpha_{p-1}}$$

$$+ \frac{1}{2\alpha_p}\left[\frac{1}{2}\log\left|\frac{\alpha_p+x}{\alpha_p-x}\right| + \operatorname{arc\,tg}\frac{x}{\alpha_p}\right]_{\infty}^{\alpha_{p+1}}$$

$$\leqq \frac{M}{p} + M\left|\frac{\log p}{p}\right| \leqq M\frac{|\log p|}{|p|},$$

the estimate being also optimal. □

Acknowledgments. The author wishes to thank the referees who suggested improvements of the paper.

REFERENCES

[1] J. M. Ball and M. Slemrod, *Nonharmonic Fourier series and the stabilization of distributed semi-linear control systems*, Comm. Pure Appl. Math., 32 (1979), pp. 555-587.

[2] G. Chen, M. Delfour, A. M. Krall, and G. Payre, *Modelling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526-546.

[3] G. Chen, S. G. Krantz, D. W. Ma, C. E. Wayne, and H. H. West, *The Euler-Bernoulli beam equation with boundary energy dissipation*, in Operator Methods for Optimal Control Problems, Sung J. Lee, ed., Marcel Dekker, New York, 1988.

[4] G. Chen, S. G. Krantz, D. L. Russell, C. E. Wayne, H. H. West, and M. P. Coleman, *Analysis, designs, and behavior of dissipative joints for coupled beams*, SIAM J. Appl. Math., 49 (1989), pp. 1665-1693.

[5] S. G. Krantz and W. H. Paulsen, *Asymptotic eigenfrequency distributions for the N-beam Euler-Bernoulli coupled beam with dissipative joints*, preprint.

[6] F. Conrad, *Stabilization of vibrating beams by a specific feedback*, COMCON Workshop on Stabilization of Flexible Structures, Montpellier, France, 1987, in Stabilization of Flexible Structures, A. V. Balakrishnan and J. P. Zolésio, eds., Optimization Software, New York, 1988, pp. 36-51.

[7] M. Delfour and M. P. Polis, *On Issues Related to Stabilization of Large Flexible Structures*, manuscript.

[8] D. Huet, *Décomposition spectrale et opérateurs*, Presses Universitaires de France, Paris, 1976.

[9] J. Leblond and J. P. Marmorat, *Stabilization of flexible structures with unbounded input and output operators*, preprint.

[10] P. Rideau, *Contrôle d'un assemblage de poutres flexibles par des capteurs-actionneurs ponctuels*, Thesis, Ecole des Mines de Paris, Sophia-Antipolis, France, 1985.

[11] F. Riesz and B. Sz. Nagy, *Leçons d'Analyse Fonctionnelle*, Gauthiers-Villars, Paris, 1968.

[12] D. L. Russell, *Linear stabilization of the linear oscillator in Hilbert space*, J. Math. Anal. Appl., 25 (1969), pp. 663-675.

[13] Y. Sakawa, *Feedback control of second order evolution equations with damping*, SIAM J. Control Optim., 22 (1984), pp. 343-361.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 438-465, March 1990

011

BILINEAR TRANSFORMATION OF INFINITE-DIMENSIONAL STATE-SPACE SYSTEMS AND BALANCED REALIZATIONS OF NONRATIONAL TRANSFER FUNCTIONS*

RAIMUND OBER† AND STEPHEN MONTGOMERY-SMITH‡

Abstract. The bilinear transform maps the open right half plane to the open unit disk and is therefore a suitable tool for carrying over results for continuous-time systems to discrete-time systems and vice versa. Corresponding state-space formulae are widely used and well understood for the case of finite-dimensional systems. In this paper infinite-dimensional generalizations of these formulae are studied for a general class of infinite-dimensional state-space systems. In particular, it is shown that reachability and observability are carried over and that the reachability and observability gramians are preserved under this transformation. Young showed that a wide class of nonrational discrete-time transfer functions admit a balanced state-space representation. It is shown that this result carries over to the continuous-time situation via the bilinear transformation.

Key words. bilinear transformation, infinite-dimensional state-space systems, balanced realizations

AMS(MOS) subject classifications. 93C20, 93B15, 93B20, 93B28

1. Introduction. Balanced realizations for finite-dimensional systems have received a great deal of attention. They were introduced as a means of performing model reduction in an easy fashion [10] and have subsequently been used in H^∞ control theory, for example, to evaluate the Hankel norm of a linear system [5], [3]. Recently, they have been used to study parametrization problems of the set of stable linear systems [11], [13].

The elegant results obtained for finite-dimensional balanced systems brought about some interest in the problem of the extension of the notion of a balanced realization to infinite-dimensional systems. Curtain and Glover [2], as well as Glover, Curtain, and Partington [6] derived continuous-time, balanced realizations for a class of systems with nuclear Hankel operator. Young [20] developed a very general realization theory for infinite-dimensional discrete-time systems.

The motivation for this paper was to show that a large class of systems that includes most H^∞ transfer functions have balanced realizations. Transfer functions in H^∞ are of particular interest since they are precisely the transfer functions of linear systems with L^2 bounded input-output operators. In particular, it is shown here that important systems such as a pure time delay, delayed systems with transfer functions of the form $G(s)\,e^{-sT}$, $G(s)$ nonstrictly proper rational, but also certain transfer functions with singularities on the imaginary axis such as $G(s)=\log(1+1/s)$ admit balanced or, more precisely, parbalanced realizations. These are examples of systems whose corresponding Hankel operator is not nuclear and hence they are not in the class of systems considered by Glover, Curtain, and Partington. The work by Glover, Curtain, and Partington [6] and Ober [12] has shown that balanced realizations can be successfully employed to perform model reduction for certain special classes of infinite-dimensional continuous-time systems. It is hoped that the realization theory for balanced systems developed here is not only of theoretical interest but is also a

* Received by the editors May 31, 1988; accepted for publication (in revised form) May 26, 1989.

† Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, United Kingdom.

‡ Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, CB2 1SB, United Kingdom. Present address, Department of Mathematics, University of Missouri at Columbia, Columbia, Missouri 65211.

step toward the development of model reduction tools for the important class of H^∞ transfer functions.

The realization problem for infinite-dimensional continuous-time systems has been studied by several authors. Shift realizations of infinite-dimensional continuous-time systems have been investigated, for example, by Fuhrmann [4] and Salamon [16]. Other approaches have been taken by Yamamoto [19] and Hegner [12]. The important role the Hankel operator plays in realization theory is well understood (see, e.g., Fuhrmann [4]). From this point of view it is interesting to note that such a connection is also very apparent in the realization of infinite-dimensional systems in terms of balanced realizations. For example, the realizability conditions on a transfer function are in terms of boundedness conditions and compactness conditions on the Hankel operator corresponding to the transfer function. But these can be expressed in terms of analytical properties of the transfer functions.

System theoretic developments often go in parallel for continuous-time and discrete-time systems. In finite-dimensional system theory it is common practice to derive results for one class of systems and then map these over to the other by using a bilinear transformation or the corresponding state space formulae. With this method it is often possible to avoid the repetition of lengthy derivations if results have already been obtained for one class of systems and similar results are needed for the other. The approach taken to the realization problem considered here is based on the same principal. The work by Young [20] contains very general realization results for discrete-time systems in terms of balanced realizations. We will carry these over to the continuous-time case using infinite-dimensional generalizations of the finite-dimensional methods. A major part of this paper is devoted to establishing infinite-dimensional generalizations of these techniques. It is shown that such generalizations are indeed possible and are especially suited to the study of observability and reachability properties, which are of central importance in linear systems theory. In particular, it is shown that these techniques carry over the observability and reachability operators in such a way that the observability (reachability) operator of a continuous-time system and the observability (reachability) operator of its corresponding discrete-time system are unitarily equivalent. It is hoped that such methods will become as useful in an infinite-dimensional setting as they have proved to be for finite-dimensional systems.

In essence, we will prove infinite-dimensional analogues of the following finite-dimensional results. If $C_n^{p,m}$ is the set of minimal asymptotically stable continuous-time systems $(A_c, B_c, C_c, D_c) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n} \times \mathbb{R}^{p\times m}$ and $D_n^{p,m}$ is the set of minimal asymptotically stable discrete-time systems $(A_d, B_d, C_d, D_d) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n} \times \mathbb{R}^{p\times m}$, then the map $T_n : D_n^{p,m} \to C_n^{p,m}$ defined by

$$T_n((A_d, B_d, C_d, D_d))$$
$$= ((I + A_d)^{-1}(A_d - I), \sqrt{2}(I + A_d)^{-1}B_d, \sqrt{2}\, C_d(I + A_d)^{-1}, D_d - C_d(I + A_d)^{-1}B_d)$$

is a bijection with inverse $T_n^{-1} : C_n^{p,n} \to D_m^{p,n}$ given by

$$T_n^{-1}((A_c, B_c, C_c, D_c))$$
$$= ((I - A_c)^{-1}(I + A_c), \sqrt{2}(I - A_c)^{-1}B_c, \sqrt{2}\, C_c(I - A_c)^{-1}, D_c + C_c(I - A_c)^{-1}B_c).$$

If $(A_c, B_c, C_c, D_c) := T_n((A_d, B_d, C_d, D_d))$, then (A_d, B_d, C_d, D_d) is a realization of the transfer function G_d, i.e., $G_d(z) = C_d(zI - A_d)^{-1}B_d + D_d$, if and only if (A_c, B_c, C_c, D_c) is a realization of the transfer function $G_c(s) := G_d((1+s)/(1-s))$, i.e., $G_c(s) = C_c(sI - A_c)^{-1}B_c + D_c$.

To deal with infinite-dimensional continuous-time systems in their full generality, it is however necessary, in contrast to discrete-time systems, to deal with unbounded input and output operators. This produces serious technical problems, and a careful setup is necessary for the definition of an infinite-dimensional continuous-time system and of the generalization of the transformation T_n.

Our approach to the definition of an infinite-dimensional system is based on the notion of a compatible system, as introduced by Helton [9]. There are, however, several differences in technical details that seem necessary to prove our result. Hedberg [8] used a form of state-space formulae to relate discrete-time shift realizations to continuous-time shift realizations. His method was later reported in the book by Fuhrmann [4]. To derive our results we had to adopt a generalization of the transformation T_n that differs from Hedberg's generalization in several respects.

In § 2 we define the objects of interest to our study, that is, infinite-dimensional discrete- and continuous-time systems. To do this it is necessary to introduce the notion of a rigged Hilbert space, as well as prove several properties of generators of semigroups that are important in our context.

We will need several results from the functional calculus for unbounded functions by Sz.-Nagy and Foias [17]. Section 3 contains a brief introduction to this functional calculus and proves propositions that we will need in later sections. The section can be skipped by readers who are not interested in detailed proofs of the main theorems of the paper.

Section 4 contains our first important results. Here we establish the transformation T relating infinite-dimensional discrete-time systems to continuous-time systems and show that it is a bijection.

State-space systems related by a unitary state-space transformation are studied in § 5. It is established that two discrete-time systems are unitarily equivalent if and only if their corresponding continuous-time systems are unitarily equivalent.

Before § 7, we need to generalize the notion of the dual of a system to infinite-dimensional systems. This is done in § 6.

Section 7 contains one of the main results of this paper. It is shown that the observability operator of a discrete-time system is unitarily equivalent to the observability operator of its corresponding continuous-time system.

Having established all the necessary tools for our treatment of infinite-dimensional state-space systems, we bring them together in § 8 where we prove a general realization result for infinite-dimensional continuous-time transfer functions in terms of balanced systems.

Great emphasis has been placed on a presentation that is as self-contained as possible. It is hoped that this paper might serve some readers as an introduction to infinite-dimensional continuous-time state-space systems.

All Hilbert spaces are assumed to be separable and defined over the complex field. The scalar product $\langle \cdot, \cdot \rangle$ is linear in the first component. The norm of a Hilbert space X is denoted by $\| \cdot \|_X$, or simply by $\| \cdot \|$. The sum of two subsets N and M of a Hilbert space X is defined by $M + N = \{x + y \mid x \in M, y \in N\}$. We denote by $(A, D(A))$ the operator A with domain of definition $D(A)$. The adjoint of the operator $(A, D(A))$ is denoted by $(A^*, D(A^*))$. The space of bounded operators from the Hilbert space X to the Hilbert space Y is denoted by $\mathscr{L}(X, Y)$, whereas $\mathscr{K}(X, Y)$ is the set of compact operators from X to Y. The symbol $\sigma_p(A)$ indicates the point spectrum of the operator A. The abbreviation RHP stands for the open right half plane. The boundary of the open unit disc $\mathbf{D}$ is denoted by $\partial\mathbf{D}$. The real part of a complex number z is denoted by $\mathrm{Re}\,(z)$.

2. Admissible discrete-time and continuous-time systems. In this section we will define the classes of discrete- and continuous-time systems that we will investigate in later parts of the paper. Whereas we can immediately state what we mean by an admissible discrete-time system, we will have to review the notion of a rigged Hilbert-space before we can give the corresponding definition of an admissible continuous-time system.

An admissible discrete-time system is defined as follows.

DEFINITION 2.1. The quadruple of operators (A_d, B_d, C_d, D_d) is called an admissible discrete-time system, with state space X, output space Y and input space U, where X, U, Y are separable Hilbert spaces, if

(i) $A_d \in \mathscr{L}(X)$ is a contraction such that $-1 \notin \sigma_p(A_d)$,
(ii) $B_d \in \mathscr{L}(U, X)$,
(iii) $C_d \in \mathscr{L}(X, Y)$,
(iv) $D_d \in \mathscr{L}(U, Y)$,
(v) A_d, B_d, C_d are such that $\lim_{\lambda \to 1, \lambda > 1} C_d(\lambda I + A_d)^{-1} B_d$ exists in the norm topology.

We write $D_X^{U,Y}$ for the set of admissible discrete-time systems with input space U, output space Y and state space X.

Remark 2.2. The technical condition (v), which is generally not very restrictive, is not necessary to define infinite-dimensional discrete-time systems. It is, however, important to study the connection between continuous-time and discrete-time systems.

We briefly introduce a number of definitions and results on strongly continuous semigroups of contractions. An excellent reference is Pazy [14].

DEFINITION 2.3. Let X be a Hilbert space. A one-parameter family $(T(t))_{t \geqq 0}$ of contractions in $\mathscr{L}(X)$ is a strongly continuous semigroup of contractions if

(i) $T(0) = I$,
(ii) $T(t+s) = T(t)T(s)$ for every $t, s \geqq 0$,
(iii) $\lim_{t \to 0} T(t)x = x$ for all $x \in X$.

The linear operator $(A, D(A))$ given by

$$Ax = \lim_{t \to 0} \frac{T(t)x - x}{t} \quad \text{for } x \in D(A), \qquad D(A) = \left\{ x \in X \;\middle|\; \lim_{t \to 0} \frac{T(t)x - x}{t} \text{ exists} \right\}$$

is called the generator of the semigroup $(T(t))_{t \geqq 0}$.

It can be shown that the generator $(A, D(A))$ uniquely determines the corresponding semigroup $(T(t))_{t \geqq 0}$. Therefore we write $T(t) =: e^{tA}$, $t \geqq 0$. We note that the generator $(A, D(A))$ is a closed linear operator whose domain $D(A)$ is dense in X. A further important property is that it is dissipative, i.e.,

$$\operatorname{Re} \langle Ax, x \rangle \leqq 0 \quad \text{for all } x \in D(A).$$

Moreover, $D(A)$ is a Hilbert space with inner product induced by the graph norm

$$\|x\|_A^2 := \|x\|_X^2 + \|Ax\|_X^2, \qquad x \in D(A).$$

Since $\|x\|_A \geqq \|x\|_X$ for $x \in D(A)$, we can embed X in $D(A)^{(\prime)}$, the set of antilinear continuous functionals on $(D(A), \|\cdot\|_A)$, by

$$E : X \to D(A)^{(\prime)}, \qquad x \mapsto (y \mapsto \langle x, y \rangle).$$

Note that $D(A)^{(\prime)}$ is a Hilbert space with norm $\|f\|' := \sup_{\|x\|_A \leqq 1} |f(x)|$. Since $\langle \cdot, \cdot \rangle$ is linear in the first component, the embedding E is linear. By the above, we have the

rigged structure

$$D(A) \subseteq X \subseteq D(A)^{(\prime)}.$$

It is well known that if $(A, D(A))$ is the generator of a strongly continuous semigroup of contractions $(e^{tA})_{t\geq 0}$ on a Hilbert space, then the adjoint $(A^*, D(A^*))$ of $(A, D(A))$ is the generator of the adjoint semigroup $(e^{tA})^*_{t\geq 0}$. Hence, we have similarly that

$$D(A^*) \subseteq X \subseteq D(A^*)^{(\prime)}.$$

If M is an operator on X such that $D(A^*) \subseteq X$ is invariant under M^*, then M can be extended to an operator $\tilde{M}$ on $D(A^*)^{(\prime)}$ by

$$\tilde{M}: D(A^*)^{(\prime)} \to D(A^*)^{(\prime)}, \qquad f(\cdot) \mapsto f(M^*(\cdot)).$$

Usually we will not distinguish between M and $\tilde{M}$ and we will write M for $\tilde{M}$.

Also, if we have a map $M: Z \to D(A^*)^{(\prime)}$, Z a Hilbert space, such that $M(Z) \subseteq X^{(\prime)} \subseteq D(A^*)^{(\prime)}$, we can consider $M: Z \to X$ using the Riesz Representation Theorem.

We are now in a position to define admissible continuous-time systems.

DEFINITION 2.4. A quadruple of operators (A_c, B_c, C_c, D_c) is called an admissible continuous-time system with state space X, input space U, and output space Y, where X, U, Y are separable Hilbert spaces, if

(i) $(A_c, D(A_c))$ is the generator of a strongly continuous semigroup of contractions on X.

(ii) $B_c: U \to (D(A_c^*)^{(\prime)}, \|\cdot\|')$ is a bounded linear operator.

(iii) $C_c: D(C_c) \to Y$ is linear with $D(C_c) = D(A_c) + (I - A_c)^{-1} B_c U$ and $C_{c|D(A_c)}: (D(A_c), \|\cdot\|_{A_c}) \to Y$ is bounded.

(iv) $C_c(I - A_c)^{-1} B_c \in \mathscr{L}(U, Y)$.

(v) A_c, B_c, C_c are such that $\lim_{s\in\mathbb{R}, s\to\infty} C_c(sI - A_c)^{-1} B_c = 0$ in the norm topology.

(vi) $D_c \in \mathscr{L}(U, Y)$.

We write $C_X^{U,Y}$ for the set of admissible continuous-time systems with input space U, output space Y, and state space X.

Before we continue to prove two lemmas that show admissible continuous-time systems are well defined, let us remark that the state space X of a system in $C_X^{U,Y}$ has the rigged structure $D(A_c) \subseteq X \subseteq D(A_c^*)^{(\prime)}$.

Remark 2.5. In Helton [9] and Fuhrmann [4] a similar definition was given for continuous-time state-space systems. There are, however, several differences between so-called compatible systems and admissible systems as defined here. Our definition of a rigged Hilbert space is slightly different from that used in Helton and Fuhrmann, where X is embedded in the dual spaces $D(A)'$ and $D(A^*)'$, rather than in the spaces of antilinear functionals $D(A)'$ and $D(A^*)'$ as adopted here. The reason for using our definition is that this naturally leads to a definition of the input operator B_c as a linear, rather than an antilinear operator. Most important, however, for the discussion later, is the imposition of (v) in our definition.

To show that the above definition is well defined, we must show that $C_c(sI - A_c)^{-1} B_c$ is well defined for all $s \in \mathbb{R}$ and that $(I - A_c)^{-1} B_c U \subseteq X$. This follows from the following two lemmas, which also contain technical results that are useful in later sections.

LEMMA 2.6. *Let $(A_c, D(A_c))$ be the generator of a strongly continuous semigroup of contractions $(e^{tA_c})_{t\geq 0}$ on the separable Hilbert space X. Then for $s \in$ RHP,*

(i) *$(sI - A_c)^{-1} X \subseteq D(A_c)$ and the map $(sI - A_c)^{-1}: (X, \|\cdot\|_X) \to (D(A_c), \|\cdot\|_{A_c})$ is bounded.*

(ii) *The map $(sI - A_c)^{-1}: (D(A_c), \|\cdot\|_{A_c}) \to (D(A_c), \|\cdot\|_{A_c})$ is bounded.*

(iii) $(sI-A_c)^{-1}D(A_c^*)^{(\prime)} \subseteq X$ *and the map* $(sI-A_c)^{-1}:(D(A_c^*)^{(\prime)}, \|\cdot\|') \to (X, \|\cdot\|_X)$ *is bounded.*

(iv) $e^{tA_c}D(A_c) \subseteq D(A_c)$ *for all* $t \in [0, \infty[$.

Proof. (i) For a proof that $(sI-A_c)^{-1}X \subseteq D(A_c)$, $s \in \mathrm{RHP}$, see Pazy [14, p. 8]. To show that $(sI-A_c)^{-1}:(X, \|\cdot\|_X) \to (D(A_c), \|\cdot\|_{A_c})$ is bounded, $s \in \mathrm{RHP}$, let $x \in X$ and consider

$$\begin{aligned}\|(sI-A_c)^{-1}x\|_{A_c}^2 &= \|(sI-A_c)^{-1}x\|_X^2 + \|A_c(sI-A_c)^{-1}x\|_X^2 \\ &\leqq \|(sI-A_c)^{-1}\|^2\|x\|_X^2 + \|s(sI-A_c)^{-1}x - (sI-A_c)(sI-A_c)^{-1}x\|_X^2 \\ &\leqq \|(sI-A_c)^{-1}\|^2\|x\|_X^2 + (|s|\,\|(sI-A_c)^{-1}\|\,\|x\|_X + \|x\|_X)^2 \\ &= (\|(sI-A_c)^{-1}\|^2 + (|s|\,\|(sI-A_c)^{-1}\|+1)^2)\|x\|_X^2,\end{aligned}$$

which proves the result.

(ii) This follows from (i) since $\|x\|_X \leqq \|x\|_{A_c}$, for $x \in D(A_c)$.

(iii) This follows by duality from (i).

(iv) See Pazy [14, p. 5]. □

By (iii) of the previous lemma and the definition of B_c we have that $(I-A_c)^{-1}B_c \subseteq X$ and so $D(C_c)$ is well defined. The following lemma shows that $C_c(sI-A_c)^{-1}B_c$ is well defined and in $\mathscr{L}(U, Y)$, for all $s \in \mathrm{RHP}$.

LEMMA 2.7. *Let* $A_c: D(A_c) \to X$ *be the generator of a strongly continuous semigroup of contractions. Let* $B_c: U \to (D(A_c^*)^{(\prime)}, \|\cdot\|')$ *be bounded and let* $C_c: D(C_c) \to Y$ *be such that* $C_{c|D(A_c)}: (D(A_c), \|\cdot\|_{A_c}) \to Y$ *is bounded, where* $D(C_c) = D(A_c) + (I-A_c)^{-1}B_cU$. *Then*

(i) $(sI-A_c)^{-1}B_cU \subseteq D(C_c)$ *for all* $s \in \mathrm{RHP}$.

(ii) *If* $C_c(I-A_c)^{-1}B_c \in \mathscr{L}(U, Y)$, *then* $C_c(sI-A_c)^{-1}B_c \in \mathscr{L}(U, Y)$ *for all* $s \in \mathrm{RHP}$.

Proof. Let $s \in \mathrm{RHP}$; then by the resolvent identity we have

$$(sI-A_c)^{-1} = (I-A_c)^{-1} + (1-s)(I-A_c)^{-1}(sI-A_c)^{-1}.$$

Since $(sI-A_c^*)^{-1}D(A_c^*) \subseteq D(A_c^*)$ we can apply B_c and obtain

$$(sI-A_c)^{-1}B_c = (I-A_c)^{-1}B_c + (1-s)(I-A_c)^{-1}(sI-A_c)^{-1}B_c.$$

Since $B_c: U \to (D(A_c^*)^{(\prime)}, \|\cdot\|)$ is bounded, it follows by Lemma 2.6(i), (iii) that

$$(I-A_c)^{-1}(sI-A_c)^{-1}B_c: U \to (D(A_c), \|\cdot\|_{A_c})$$

is continuous. This implies in particular (i), since

$$(sI-A_c)^{-1}B_cU = (I-A_c)^{-1}B_cU + (1-s)(I-A_c)^{-1}(sI-A_c)^{-1}B_cU \subseteq D(C_c).$$

Since $C_{c|D(A_c)}: (D(A_c), \|\cdot\|_{A_c}) \to Y$ is bounded and hence

$$C_c(I-A_c)^{-1}(sI-A_c)^{-1}B_c \in \mathscr{L}(U, Y),$$

we have that

$$C_c(sI-A_c)^{-1}B_c = C_c(I-A_c)^{-1}B_c + (1-s)C_c(I-A_c)^{-1}(sI-A_c)^{-1}B_c \in \mathscr{L}(U, Y). \quad \square$$

Remark 2.8. It is useful to note that using the identification of $X^{(\prime)}$ and X via the Riesz Representation Theorem we have that for $u \in U$ the functional $(sI-A_c)^{-1}B_c(u): D(A_c^*) \to \mathbb{C}$ is given by

$$x \mapsto (sI-A_c)^{-1}B_c(u)[x] = B_c(u)[(\bar{s}I-A_c^*)^{-1}x] = \langle (sI-A_c)^{-1}B_cu, x\rangle.$$

3. The functional calculus by Sz.-Nagy–Foias. In this section we will review some results on the functional calculus by Sz.-Nagy-Foias and prove two technical results that are fundamental to the main results of this paper. This section is, however, only necessary for an understanding of the proofs of some of the theorems presented in later sections. Those theorems themselves are largely formulated without reference to the functional calculus discussed here.

Since we do not assume that the reader is fully familiar with the functional calculus as developed in Sz.-Nagy and Foias [17] we give a brief summary of those results necessary for our applications.

We first consider a standard result of functional calculus. Let $\mathscr{A}$ be the set of functions given by

$$a(z)=\sum_{k=0}^{\infty} c_k z^k \quad \text{such that} \quad \sum_{k=0}^{\infty} |c_k|<\infty.$$

Then $\mathscr{A}$ is an algebra with the involution $a\mapsto a^*$ given by $a^*(z):=\overline{a(\bar{z})}$. Note that a function in $\mathscr{A}$ is analytic on $\mathbf{D}$ and continuous on $\bar{\mathbf{D}}$.

For a contraction T on a Hilbert space X, we define $a(T)=\sum_{k=0}^{\infty} c_k T^k$. The sum converges in the operator norm and hence the operator $a(T)$ is well defined. The following theorem states the fundamental result concerning the functional calculus for functions in $\mathscr{A}$.

THEOREM 3.1. *For a contraction T on a Hilbert space X, the map*

$$\mathscr{A}\to\mathscr{L}(X), \qquad a(z)\mapsto a(T)$$

is an algebra homomorphism. In particular, $a(T)b(T)=b(T)a(T)$, for $a, b\in\mathscr{A}$.

The functions that are important in our context are:

(a) $\phi: z\to(z-1)/(z+1)$,

(b) $\varphi_t: z\mapsto e^{t((z-1)/(z+1))}$, $t\geqq 0$,

(c) $\mu: z\mapsto 1/(1+z)$,

(d) $\delta_t: z\mapsto 1/(1+z)\, e^{t((z-1)/(z+1))}$, $t\geqq 0$.

None of these functions are in $\mathscr{A}$ and hence we must consider extensions of the functional calculus of Theorem 3.1. Note, however, that the functions $z\mapsto\phi(rz)$, $z\mapsto\varphi_t(rz)$, $z\mapsto\mu(rz)$, and $z\mapsto\delta_t(rz)$, $0<r<1$ are in $\mathscr{A}$.

Next we exploit the observation that for each function $u\in H^\infty$ the function $z\mapsto u(rz)$, $0<r<1$, is in $\mathscr{A}$ and discuss functions for which the limit $\lim_{r\to 1-0} u(rT)$ is well defined in the following sense.

DEFINITION 3.2. Let T be a contraction on X. H_T^∞ is the set of those functions $u\in H^\infty$ such that

$$u(T):=\lim_{r\to 1-0} u(rT)$$

exists in the strong operator topology.

Before we can describe a subset of H_T^∞, we must consider contractions in some detail. A subspace Y of a Hilbert space X is called reducing for $T\in\mathscr{L}(X)$ if T maps Y onto itself. A contraction T in $\mathscr{L}(X)$ is called completely nonunitary if there is no nonzero reducing subspace Y of X such that $T_{|Y}$ is unitary. To every contraction T on the space X there corresponds a decomposition $X=X_1\oplus X_2$ into an orthogonal sum of two subspaces X_1 and X_2 reducing T such that $T_1:=T_{|X_1}$ is unitary and $T_2:=T_{|X_2}$ is completely nonunitary. The canonical decomposition of T is denoted by $T=T_1\oplus T_2$. Recall that each unitary operator U has a spectral decomposition $U=\int_0^{2\pi} e^{it}\, dE_t$ for some spectral family $\{E_t\}_{0\leqq t\leqq 2\pi}$ and spectral measure E_U on the unit circle.

THEOREM 3.3. *H_T^∞ contains the functions $u \in H^\infty$ for which the set*

$$C_u = \{z \in \partial \mathbf{D} \mid u(z) \text{ has no nontangential limit at } z\}$$

has measure zero with respect to the spectral measure E_{T_1} corresponding to the unitary part T_1 of T.

Remark 3.4. Now we consider the functions φ_t, $t \geqq 0$, as defined in (b). We clearly have that $\varphi_t \in H^\infty$ and $C_{\varphi_t} \subseteq \{-1\}$. For a contraction T such that -1 is not an eigenvalue of T, -1 is also not an eigenvalue of the unitary part T_1 of T and hence $E_{T_1}(\{-1\}) = 0$, which shows that $\varphi_t \in H_T^\infty$, $t \geqq 0$.

We will not explore the properties of H_T^∞ in general, but consider the special case of the functions φ_t, $t \geqq 0$. These are of importance in connection with semigroup theory. Before we can state the next theorem establishing this role, we need to introduce some additional notation. Let A_c be the generator of a strongly continuous semigroup of contractions $(e^{tA_c})_{t \geqq 0}$; then

$$A_d = (I + A_c)(I - A_c)^{-1}$$

is called the cogenerator of the semigroup $(e^{tA_c})_{t \geqq 0}$ that can be shown to be a contraction such that -1 is not an eigenvalue of A_d. The generator A_c can be expressed by A_d as

$$A_c = (I + A_d)^{-1}(A_d - I).$$

The following theorem states that if given a contraction T such that -1 is not an eigenvalue of T, then $(\varphi_t(T))_{t \geqq 0}$ is a semigroup of contractions with generator $(I+T)^{-1}(T-I)$ and cogenerator T.

THEOREM 3.5. *Let T be a contraction on X. In order that there exists a strongly continuous semigroup of contractions $(T(t))_{t \geqq 0}$ whose cogenerator equals T, it is necessary and sufficient that -1 is not an eigenvalue of T. If this is the case, then $(T(t))_{t \geqq 0}$ is determined by*

$$T(t) = \varphi_t(T), \qquad t \geqq 0$$

with generator $A_c = (I+T)^{-1}(T-I)$.

Proof. The proof follows from Sz.-Nagy and Foias [17, p. 142], replacing T by $-T$. □

We will now consider unbounded functions in order to deal with ϕ, μ, and δ_t. If $T \in \mathcal{L}(X)$ is a contraction such that $-1 \notin \sigma_p(T)$, then it is easily checked that ϕ, μ, and δ_t, $t \geqq 0$ are in the set of functions N_T defined as follows.

DEFINITION 3.6. For a contraction T in $\mathcal{L}(X)$, denote by K_T^∞ the class of functions $v \in H_T^\infty$ for which $v(T)^{-1}$ exists and has dense domain. Let N_T be the class of functions w that admit a representation

$$w = \frac{u}{v}, \quad u \in H_T^\infty, \quad v \in K_T^\infty.$$

For $w \in N_T$, we define $w(T) = v(T)^{-1}u(T)$.

The following proposition states that for a certain subset of N_T we, in fact, have the commutativity property $w(T) = v(T)^{-1}u(T) = u(T)v(T)^{-1}$.

PROPOSITION 3.7. *Let u, v be continuous on $\bar{\mathbf{D}}$, analytic on $\mathbf{D}$ and have no common zeros in $\bar{\mathbf{D}}$. If $v \in K_T^\infty$ for a contraction T, then*

$$v(T)^{-1}u(T) = u(T)v(T)^{-1}.$$

Remark 3.8. Let T be a contraction T such that $-1 \notin \sigma_p(T)$; then, applying the previous proposition to ϕ, we obtain $(I-T)(I+T)^{-1} = (I+T)^{-1}(I-T)$.

The following theorem provides us with techniques to deal with functions in N_T.

THEOREM 3.9. (i) *Let T be a contraction in $\mathscr{L}(X)$ and let $w \in N_T$ be analytic on* $\mathbf{D}$. *If for $x \in X$ we have that*

$$\sup_{0<r<1} \|w(rT)x\| < \infty,$$

it follows that $x \in D(w(T))$ and

$$w(rT)x \to w(T)x$$

weakly as $r \to 1-0$.

(ii) *Suppose the functions u, v are continuous on $\bar{\mathbf{D}}$, analytic on $\mathbf{D}$, and have no common zeros in $\bar{\mathbf{D}}$. We assume that v has no zeros in $\mathbf{D}$ and that it does not vanish on $\partial\mathbf{D}$ except at points of measure zero with respect to the spectral measure E_{T_1} of the unitary part T_1 of T. Moreover, we assume that there exists a constant M such that $|v(\lambda)/v(r\lambda)| \leqq M$ for $\lambda \in \mathbf{D}$, $0<r<1$. Then $w = u/v$ belongs to the class N_T and is analytic in $\mathbf{D}$.*

The condition

$$\sup_{0<r<1} \|w(rT)x\| < \infty$$

characterizes the vectors in $D(w(T))$.

For each $x \in D(w(T))$,

$$w(rT)x \to w(T)x$$

strongly as $r \to 1-0$.

Having reviewed the functional calculus by Sz.-Nagy–Foias, we are now in a position to prove two results that will be a key to results in later sections. Whereas the first proposition deals with the function μ, the second proposition establishes properties of the function δ_t.

PROPOSITION 3.10. *Let T be a contraction on X such that -1 is not an eigenvalue of T. Then*

$$\lim_{\substack{\lambda \to 1 \\ \lambda > 1}} (\lambda I + T)^{-1}x = (I+T)^{-1}x,$$

for $x \in D((I+T)^{-1})$.

Moreover, $x \in D((I+T)^{-1})$ if and only if $\sup_{0<r<1} \|(I+rT)^{-1}x\| < \infty$.

Proof. Let $u = 1$ and $v = 1+z$, so v only vanishes at $z=-1$. Since -1 is not an eigenvalue of T we have that $E_{T_1}(\{-1\}) = 0$. Using $w := u/v$ and $r := 1/\lambda$, the result now follows from Theorem 3.9(ii). □

PROPOSITION 3.11. *Let T be a contraction on X such that -1 is not an eigenvalue of T. If $x \in D((I+T)^{-1})$, then for all $t \geqq 0$,*

(1) $\sup_{0<r<1} \|(I+rT)^{-1} e^{t(rT+I)^{-1}(rT-I)}x\| \leqq \sup_{0<r<1} \|(I+rT)^{-1}x\| < \infty$,

(2) $(I+rT)^{-1} e^{t(rT+I)^{-1}(rT-I)}x \to (I+T)^{-1} e^{t(T+I)^{-1}(T-I)}x$ *weakly as* $r \to 1-0$.

Proof. Write for $\delta_t(z) = 1/(1+z)\, e^{t((z-1)/(z+1))} = \mu(z)\varphi_t(z)$, with $\mu(z) = 1/(z+1)$ and $\varphi_t(z) = e^{t((z-1)/(z+1))}$, $t \geqq 0$. Then we have that $\delta_t \in N_T$, $t \geqq 0$, since $\mu^{-1} \in K_T^\infty$ and since $\varphi_t \in H_T^\infty$, $t \geqq 0$, by Remark 3.4.

As $\varphi_t(rz), \mu(rz) \in \mathscr{A}$, for $0<r<1$, $t \geqq 0$, we have that $\varphi_t(rT)\mu(rT) = \mu(rT)\varphi_t(rT)$. Also note that $\varphi_t(rT)$ is a contraction by Theorem 3.5. Hence we obtain for $t \geqq 0$ and

$x \in D(\mu(T)) = D((I+T)^{-1})$ that

$$\begin{aligned}\sup_{0<r<1} \|\delta_t(rT)x\| &= \sup_{0<r<1} \|\mu(rT)\varphi_t(rT)x\| = \sup_{0<r<1} \|\varphi_t(rT)\mu(rT)x\| \\ &\leqq \sup_{0<r<1} \|\varphi_t(rT)\| \, \|\mu(rT)x\| \leqq \sup_{0<r<1} \|(I+rT)^{-1}x\| \\ &< \infty\end{aligned}$$

where the last inequality follows from Proposition 3.10. The chain of inequalities shows (1).

Thus we have by Theorem 3.9(i) that $D((I+T)^{-1}) \subseteq D(\delta_t(T))$ and that for $x \in D((I+T)^{-1})$ we have,

$$\varphi_t(rT)x \to \varphi_t(T)x$$

weakly as $r \to 1-0$, which proves (2). □

4. A transformation between discrete- and continuous-time systems. We will now introduce a transformation T relating systems in $D_X^{U,Y}$ to systems in $C_X^{U,Y}$ and vice versa. This transformation, which is inspired by a bilinear transformation mapping the unit disk to the right-half plane, is often used for finite-dimensional systems to carry over results from discrete-time systems to continuous-time systems (see, e.g., Glover [5], Ober [11]. Hedberg [8] and Fuhrmann [4] used this approach to prove the existence of state-space realizations for continuous-time systems with transfer function in a certain class of H^∞ functions. The same idea is used here, the specific definitions are, however, somewhat different to avoid certain technical problems.

We first consider the map $T: D_X^{U,Y} \to C_X^{U,Y}$.

THEOREM 4.1. *Let* $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$; *then* $T((A_d, B_d, C_d, D_d)) := (A_c, B_c, C_c, D_c) \in C_X^{U,Y}$, *where*

(i) $A_c := (I+A_d)^{-1}(A_d - I) = (A_d - I)(I+A_d)^{-1}$, $D(A_c) := D((I+A_d)^{-1})$, *and* A_c *generates a strongly continuous semigroup of contractions on* X *given by* $\varphi_t(A_d)$, $t \geqq 0$, *with* $\varphi_t(z) = e^{t((z-1)/(z+1))}$.

(ii) $B_c := \sqrt{2}(I+A_d)^{-1}B_d : U \to D(A_c^*)'$,

$$u \mapsto \sqrt{2}(I+A_d)^{-1}B_d(u)[\,\cdot\,] := \sqrt{2}\langle B_d(u), (I+A_d^*)^{-1}(\,\cdot\,)\rangle_X.$$

(iii) $C_c : D(C_c) \to Y$, $\quad x \mapsto \lim\limits_{\substack{\lambda \to 1 \\ \lambda > 1}} \sqrt{2}\, C_d(\lambda I + A_d)^{-1}x$,

where $D(C_c) = D(A_c) + (I - A_c)^{-1}B_cU$. *On* $D(A_c)$ *we have,*

$$C_{c|D(A_c)} = \sqrt{2}\, C_d(I+A_d)^{-1}.$$

(iv) $D_c := D_d - \lim_{\lambda \to 1, \lambda > 1} C_d(\lambda I + A_d)^{-1}B_d$.

Moreover, let the admissible discrete-time system (A_d, B_d, C_d, D_d) *be a realization of the transfer function*

$$G_d(z) : \mathbb{C} \setminus \bar{\mathbf{D}} \to \mathscr{L}(U, Y),$$

i.e., $G_d(z) = C_d(zI - A_d)^{-1}B_d + D_d$ *for* $z \in \mathbb{C} \setminus \bar{\mathbf{D}}$.

Then, $(A_c, B_c, C_c, D_c) = T((A_d, B_d, C_d, D_d))$ *is an admissible continuous-time realization of the transfer function*

$$G_c(s) := G_d\left(\frac{1+s}{1-s}\right) : \mathrm{RHP} \to \mathscr{L}(U, Y).$$

Proof. We must check that conditions (i)-(vi) of Definition 2.4 are satisfied.

(i) This follows from Theorem 3.5. The fact that $A_c = (A_d - I)(I + A_d)^{-1} = (I + A_d)^{-1}(A_d - I)$ was shown in Remark 3.8.

(ii) Let $u \in U$, $x \in D(A^*)$. Then, since $\frac{1}{2}(I - A_c) = (I + A_d)^{-1}$,

$$\begin{aligned} |B_c(u)[x]| &= |\sqrt{2}\langle B_d(u), (I + A_d^*)^{-1}[x]\rangle_X| \\ &\leqq \frac{1}{\sqrt{2}} \|B_d(u)\|_X \|(I - A_c^*)x\|_X \\ &\leqq \|B_d\|_{\mathscr{L}(U,X)} \|u\|_U (\|x\|_X^2 + \|A_c^* x\|_X^2)^{1/2}. \end{aligned}$$

This implies that $B_c(u) \in D(A_c^*)^{(\prime)}$ and that $B_c: U \to D(A_c^*)^{(\prime)}$ is continuous.

(iii) We first note that, by Proposition 3.10, C_c is defined on $D(A_c) = D((I + A_d)^{-1})$, and that $C_{c|D(A_c)} = \sqrt{2}\, C_d(I + A_d)^{-1}$.

To show that $C_{c|D(A_c)}$ is continuous with respect $\|\cdot\|_{A_c}$, we see that for $x \in D(A_c)$, we have

$$\begin{aligned} \|C_c x\|_Y &= \frac{1}{\sqrt{2}} \|C_d(I - A_c)x\|_Y \\ &\leqq \|C_d\|_{\mathscr{L}(X,Y)} (\|x\|_X^2 + \|A_c x\|_X^2)^{1/2}. \end{aligned}$$

It remains to show that $\lim_{\lambda \to 1, \lambda > 1} C_d(\lambda I + A_d)^{-1} x$ exists for $x \in (I - A_c)^{-1} B_c U$. First note that $(I - A_c)^{-1} B_c = (1/\sqrt{2}) B_d$, for if $x \in D(A_c^*)$, $u \in U$, then

$$\begin{aligned} (I - A_c)^{-1} B_c(u)[x] &= B_c(u)[(I - A_c^*)^{-1} x] \\ &= \sqrt{2}\, \langle B_d(u), (I + A_d^*)^{-1}(I - A_c^*)^{-1} x\rangle \\ &= \frac{1}{\sqrt{2}} \langle B_d(u), x\rangle \end{aligned}$$

where we have used the identity $(I - A_c^*)^{-1} = \frac{1}{2}(I + A_d^*)$. Now we see that

$$\lim_{\substack{\lambda \to 1 \\ \lambda > 1}} \sqrt{2}\, C_d(\lambda I + A_d)^{-1}(I - A_c)^{-1} B_c u = \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} C_d(\lambda I + A_d)^{-1} B_d u$$

exists by the admissibility of (A_d, B_d, C_d, D_d).

(iv) We must show that $C_c(I - A_c)^{-1} B_c \in \mathscr{L}(U, Y)$. But by the proof of (iii), we know that $(I - A_c)^{-1} B_c = (1/\sqrt{2}) B_d$, and hence that

$$C_c(I - A_c)^{-1} B_c = \frac{1}{\sqrt{2}} C_c B_d = \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} C_d(\lambda I + A_d)^{-1} B_d \in \mathscr{L}(U, Y)$$

by the admissibility of (A_d, B_d, C_d, D_d).

(v) This will be shown after the remaining parts of the theorem have been proved.

(vi) The boundedness of D_c follows since $D_d \in \mathscr{L}(U, Y)$ and $\lim_{\lambda \to 1, \lambda > 1} C_d(\lambda I + A_d)^{-1} B_d \in \mathscr{L}(U, Y)$.

We will now prove the statements on the transformation of transfer functions. We have for $s \in \mathrm{RHP}$ that $(1 + s)/(1 - s) \in \mathbb{C} \setminus \bar{\mathbf{D}}$ and hence,

$$\begin{aligned} G_c(s) = G_d\left(\frac{1+s}{1-s}\right) &= C_d\left(\left(\frac{1+s}{1-s}\right) I - A_d\right)^{-1} + D_d \\ &= (1 - s) C_d((I - A_d) + s(I + A_d))^{-1} B_d + D_d \\ &= (1 - s) C_d(I + A_d)^{-1}(sI - (A_d - I)(I + A_d)^{-1})^{-1} B_d + D_d. \end{aligned}$$

The last identity is well defined, since

$$(sI-(A_d-I)(I+A_d)^{-1})^{-1}B_dU=(sI-A_c)^{-1}B_dU\subseteq D(A_c)=D((I+A_d)^{-1}).$$

Hence, $G_c(s)=(1-s)(1/\sqrt{2})C_c(sI-A_c)^{-1}B_d+D_d$.

Now if we extend the range of $(sI-A_c)^{-1}B_d$ to $D(A_c^*)^{(\prime)}$, then we can show that $(sI-A_c)^{-1}B_d=\sqrt{2}/(1-s)(sI-A_c)^{-1}B_c-1/(1-s)B_d$. For if $x\in D(A_c^*)$, then we have, using the resolvent identity, that

$$\begin{aligned}
&\langle (sI-A_c)^{-1}B_d(u),x\rangle_X\\
&\quad=\langle B_d(u),(I-A_c^*)(I-A_c^*)^{-1}(\bar{s}I-A_c^*)^{-1}x\rangle_X\\
&\quad=\left\langle B_d(u),(I-A_c^*)\frac{1}{(1-\bar{s})}[(\bar{s}I-A_c^*)^{-1}-(I-A_c^*)^{-1})]x\right\rangle_X\\
&\quad=\frac{1}{1-s}(\langle B_d(u),(I-A_c^*)(\bar{s}I-A_c^*)^{-1}x\rangle_X-\langle B_d(u),x\rangle_X)\\
&\quad=\frac{\sqrt{2}}{1-s}[(sI-A_c)^{-1}B_c(u)](x)-\frac{1}{1-s}\langle B_d(u),x\rangle_X.
\end{aligned}$$

But we know that $B_dU\subseteq D(C_c)$ and $(sI-A_c)^{-1}B_cU\subseteq D(C_c)$ for $s\in$ RHP. Hence

$$\begin{aligned}
G_c(s)&=(1-s)\frac{1}{\sqrt{2}}C_c(sI-A_c)^{-1}B_d+D_d\\
&=C_c(sI-A_c)^{-1}B_c-\frac{1}{\sqrt{2}}C_cB_d+D_d\\
&=C_c(sI-A_c)^{-1}B_c-\lim_{\substack{\lambda\to 1\\ \lambda>1}}C_d(\lambda I+A_d)^{-1}B_d+D_d\\
&=C_c(sI-A_c)^{-1}B_c+D_c,
\end{aligned}$$

and so (A_c, B_c, C_c, D_c) is a state-space realization of $G_c(s)$.

To finish the proof, it remains to show (v) of Definition 2.4. By the admissibility of (A_d, B_d, C_d, D_d) we obtain

$$\begin{aligned}
\lim_{\substack{s\in\mathbb{R}\\ s\to\infty}}C_c(sI-A_c)^{-1}B_c&=\lim_{\substack{s\in\mathbb{R}\\ s\to\infty}}G_c(s)-D_c\\
&=\lim_{\substack{s\in\mathbb{R}\\ s\to\infty}}C_d\left(\left(\frac{1+s}{1-s}\right)I-A_d\right)^{-1}B_d+D_d-D_c\\
&=-\lim_{\substack{\lambda\to 1\\ \lambda>1}}C_d(\lambda I+A_d)^{-1}B_d+D_d-D_c\\
&=0,
\end{aligned}$$

which completes the proof. □

Before we consider the map $T^{-1}: C_X^{U,Y}\to D_X^{U,Y}$ we need the following lemma, which gives a version of the resolvent identity for not necessarily bounded resolvents.

LEMMA 4.2. *Let $A_d: X\to X$ be a contraction such that $-1\notin\sigma_p(A_d)$. Then for $z\in\mathbb{C}$, such that $|z|>1$ and for $x\in D((I+A_d^*)^{-1})$, we have*

$$(\bar{z}+1)(I+A_d^*)^{-1}(\bar{z}I-A_d^*)^{-1}x=(I+A_d^*)^{-1}x+(\bar{z}I-A_d^*)^{-1}x.$$

Proof. We first must show that if $x \in D((I+A_d^*)^{-1})$, then $(\bar{z}I - A_d^*)^{-1}x \in D((I+A_d^*)^{-1})$. We know by Theorem 4.1 that $A_c = (I+A_d)^{-1}(A_d - I)$ is the generator of a strongly continuous semigroup of contractions, such that $D(A_c^*) = D((I+A_d^*)^{-1})$.

Since $|z| > 1$, we have that $s = (z-1)/(z+1) \in \text{RHP}$. Hence $(\bar{s}I - A_c^*)^{-1}$ is bounded. But

$$(\bar{s}I - A_c^*)^{-1} = (\bar{z}+1)(\bar{z}(I - A_c^*) - (I + A_c^*))^{-1} = (\bar{z}+1)(\bar{z}I - A_d^*)^{-1}(I - A_c^*)^{-1}.$$

Thus $(\bar{s}I - A_c^*)^{-1}(I - A_c^*) = (\bar{z}+1)(\bar{z}I - A_d^*)^{-1}$ and hence, since $(\bar{s}I - A_c^*)^{-1}X \subseteq D(A_c^*)$ by Lemma 2.6, we have that

$$\begin{aligned}(\bar{z}+1)(\bar{z}I - A_d^*)^{-1}D(A_c^*) &= (\bar{s}I - A_c^*)^{-1}(I - A_c^*)D(A_c^*) \\ &\subseteq (\bar{s}I - A_c^*)^{-1}X \subseteq D(A_c^*) = D((I+A_d^*)^{-1}),\end{aligned}$$

which shows the claim.

To prove the statement of the lemma, let $y := (\bar{z}I - A_d^*)^{-1}x \in D((I+A_d^*)^{-1})$. Then,

$$(\bar{z}+1)y = (\bar{z}I - A_d^*)y + (I + A_d^*)y.$$

Since $y \in D((I+A_d^*)^{-1})$, we can apply $(I+A_d^*)^{-1}$ from the left to obtain

$$(\bar{z}+1)(I+A_d^*)^{-1}y = (I+A_d^*)^{-1}(\bar{z}I - A_d^*)y + y,$$

and hence $(\bar{z}+1)(I+A_d^*)^{-1}(\bar{z}I - A_d^*)^{-1}x = (I+A_d^*)^{-1}x + (\bar{z}I - A_d^*)^{-1}x$. □

THEOREM 4.3. *Let* $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$; *then* $T^{-1}((A_c, B_c, C_c, D_c)) := (A_d, B_d, C_d, D_d) \in D_X^{U,Y}$, *where*

(i) $A_d := (I+A_c)(I-A_c)^{-1}$, *and for* $x \in D(A_c)$ *we have that* $A_d x = (I-A_c)^{-1}(I+A_c)x$.

(ii) $B_d := \sqrt{2}(I-A_c)^{-1}B_c$.

(iii) $C_d := \sqrt{2}\, C_c(I-A_c)^{-1}$.

(iv) $D_d := C_c(I-A_c)^{-1}B_c + D_c$.

Moreover, let the admissible continuous-time system (A_c, B_c, C_c, D_c) *be a realization of the transfer function*

$$G_c(s) : \text{RHP} \to \mathscr{L}(U, Y),$$

i.e., $G_c(s) = C_c(sI - A_c)^{-1}B_c + D_c$ *for* $s \in \text{RHP}$.

Then, $(A_d, B_d, C_d, D_d) = T^{-1}((A_c, B_c, C_c, D_c))$ *is an admissible discrete-time realization of the transfer function*

$$G_d(z) := G_c\left(\frac{z-1}{z+1}\right) : \mathbb{C} \setminus \bar{\mathbf{D}} \to \mathscr{L}(\mathbf{U}, \mathbf{Y}).$$

Proof. We must show that (A_d, B_d, C_d, D_d) satisfies conditions (i)-(v) of Definition 2.1.

(i) Let $x \in X$ and define $y = (I-A_c)^{-1}x \in D(A_c)$; then

$$\begin{aligned}\|A_d x\|^2 = \|(I+A_c)y\|^2 &= \langle y, y\rangle + \langle A_c y, A_c y\rangle + 2\,\text{Re}\,\langle A_c y, y\rangle \\ &= \langle (I-A_c)y, (I-A_c)y\rangle + 4\,\text{Re}\,\langle A_c y, y\rangle = \|x\|^2 + 4\,\text{Re}\,\langle A_c y, y\rangle \\ &\leqq \|x\|^2\end{aligned}$$

since $\text{Re}\,\langle A_c y, y\rangle \leqq 0$ as A_c is dissipative, being the generator of a strongly continuous semigroup of contractions. This shows that A_d is a contraction.

It is easily verified that $(I+A_c)(I-A_c)^{-1}x=(I-A_c)^{-1}(I+A_c)x$, $x\in D(A_c)$, as claimed in the theorem and that $-1\notin\sigma_p(A_d)$.

(ii) This follows in a straightforward way from Lemma 2.6.

(iii) Since

$$C_{c|D(A_c)}:(D(A_c),\|\cdot\|_{A_c})\to Y$$

and

$$(I-A_c)^{-1}:(X,\|\cdot\|_X)\to(D(A_c),\|\cdot\|_{A_c})$$

are continuous, we have that

$$C_d=\sqrt{2}\,C_c(I-A_c)^{-1}:(X,\|\cdot\|_X)\to Y$$

is continuous.

(iv) Since by assumption $C_c(I-A_c)^{-1}B_c\in\mathscr{L}(U,Y)$ and $D_c\in\mathscr{L}(U,Y)$, we have that $D_d\in\mathscr{L}(U,Y)$.

(v) Before we prove (v) we first show the last statement of the theorem.

Let $z\in\mathbb{C}$, such that $|z|>1$; then $s=(z-1)/(z+1)\in$ RHP. By definition

$$G_d(z)=C_c\left(\frac{z-1}{z+1}I-A_c\right)^{-1}B_c+D_c.$$

Consider $((z-1)/(z+1)I-A_c)^{-1}B_c$. Let $u\in U$, $x\in D(A_c^*)$; then

$$\begin{aligned}\left(\frac{z-1}{z+1}I-A_c\right)^{-1}B_c(u)[x]&=B_c(u)\left[\left(\frac{\bar z-1}{\bar z+1}I-A_c^*\right)^{-1}x\right]\\&=(z+1)B_c(u)[(\bar z(I-A_c^*)-(I+A_c^*))^{-1}x]\\&=(z+1)B_c(u)[(\bar zI-A_d^*)^{-1}(I-A_c^*)^{-1}x].\end{aligned}$$

But using the fact that $(I-A_c^*)=2(I+A_d^*)^{-1}$ we obtain,

$$\begin{aligned}B_c(u)[x]&=\langle(I-A_c)^{-1}B_c(u),(I-A_c^*)x\rangle\\&=\sqrt{2}\langle B_d(u),(I-A_d^*)^{-1}x\rangle.\end{aligned}$$

Hence

$$\begin{aligned}&\left(\frac{z-1}{z+1}I-A_c\right)^{-1}B_c(u)[x]\\&\quad=(z+1)B_c(u)[(\bar zI-A_d^*)^{-1}(I-A_c^*)^{-1}x]\\&\quad=\sqrt{2}(z+1)\langle B_d(u),(I+A_d^*)^{-1}(\bar zI-A_d^*)^{-1}(I-A_c^*)^{-1}x\rangle\\&\quad=\sqrt{2}\langle B_d(u),(\bar zI-A_d^*)^{-1}(I-A_c^*)^{-1}x\rangle\\&\qquad+\sqrt{2}\langle B_d(u),(I+A_d^*)^{-1}(I-A_c^*)^{-1}x\rangle\\&\quad=\sqrt{2}\langle B_d(u),(\bar zI-A_d^*)^{-1}(I-A_c^*)^{-1}x\rangle+\frac{1}{\sqrt{2}}\langle B_d(u),x\rangle\end{aligned}$$

where the second but last equation uses Lemma 4.2, noting that $(I-A_c^*)^{-1}x\in D(A_c^*)=D((I+A_d^*)^{-1})$. Thus

$$\left(\frac{z-1}{z+1}I-A_c\right)^{-1}B_c(u)=\sqrt{2}(I-A_c)^{-1}(zI-A_d)^{-1}B_d(u)+\frac{1}{\sqrt{2}}B_d(u)\in X.$$

Note that by Lemma 2.6 $(I-A_c)^{-1}X \subseteq D(A_c) \subseteq D(C_c)$ and hence

$$\sqrt{2}(I-A_c)^{-1}(zI-A_d)^{-1}B_dU \subseteq D(C_c).$$

Since $B_dU \subset D(C_c)$ we can apply C_c, and we obtain

$$\begin{aligned} C_c\left(\frac{z-1}{z+1}I-A_c\right)^{-1}B_c+D_c &= \sqrt{2}\,C_c(I-A_c)^{-1}(zI-A_d)^{-1}B_d+\frac{1}{\sqrt{2}}C_cB_d+D_c \\ &= C_d(zI-A_d)^{-1}B_d+C_c(I-A_c)^{-1}B_c+D_c \\ &= C_d(zI-A_d)^{-1}B_d+D_d. \end{aligned}$$

Thus (A_d, B_d, C_d, D_d) is a realization of $G_d(z)$.

We are now in a position to prove (v) of Definition 2.1, i.e., that $\lim_{\lambda>1,\lambda\to 1} C_d(\lambda I+A_d)^{-1}B_d$ exists in the norm topology. By the admissibility of (A_c, B_c, C_c, D_c), we have that

$$\begin{aligned} \lim_{\substack{\lambda\to 1\\ \lambda>1}} C_d(\lambda I+A_d)^{-1}B_d &= -\lim_{\substack{\mu<-1\\ \mu\to -1}} G_d(\mu)+D_d = -\lim_{\substack{\mu<-1\\ \mu\to -1}} G_c\left(\frac{\mu-1}{\mu+1}\right)+D_d \\ &= C_c(I-A_c)^{-1}B_c, \end{aligned}$$

which implies the result. □

Combining the previous two theorems, we can show the following corollary, whose proof is straightforward.

COROLLARY 4.4. *The map* $T: D_X^{U,Y} \to C_X^{U,Y}$ *is a bijection with inverse* $T^{-1}: C_X^{U,Y} \to D_X^{U,Y}$.

Remark 4.5. The following identities that have been used in the above proofs are worthwhile noting for later use:

$$\tfrac{1}{2}(I-A_c)=(I+A_d)^{-1} \quad \text{and} \quad (I-A_c)^{-1}=\tfrac{1}{2}(I+A_d).$$

5. Unitary state-space transformations. In this section we will discuss briefly the effect of a unitary transformation $V: X_1 \to X_2$ of the state space on state-space systems. This discussion will be important in § 8 where we will show that a (par-) balanced realization is unique up to a unitary state-space transformation. The first two propositions show that such an operation is well defined and does not change the transfer function. The last result shows that unitarily equivalent systems are carried over by the map $T: D_X^{U,Y} \to C_X^{U,Y}$ and its inverse.

We first consider unitary state-space transformations for admissible discrete-time systems.

PROPOSITION 5.1. *Let* $(A_d, B_d, C_d, D_d) \in D_{X_1}^{U,Y}$. *If* X_2 *is another Hilbert space and* $V: X_1 \to X_2$ *is a unitary operator, then*

(1) $(VA_dV^*, VB_d, C_dV^*, D_d) \in D_{X_2}^{U,Y}$.

(2) If (A_d, B_d, C_d, D_d) *is a state space realization of the transfer function*

$$G_d(s): \mathbb{C}\backslash\bar{\mathbf{D}} \to L(U, Y),$$

then the $(VA_dV^*, VB_d, C_dV^*, D_d)$ *is a state-space realization of the same transfer function.*

Proof. The proof is straightforward. □

The following proposition, whose proof is straightforward, gives the analogous result for continuous-time systems.

PROPOSITION 5.2. *Let* $((A_c, D(A_c)), B_c, C_c, D_c) \in C_{X_1}^{U,Y}$. *If* X_2 *is another Hilbert space and* $V: X_1 \to X_2$ *is a unitary operator, then*

(1) $((VA_cV^*, VD(A_c)), VB_c, (C_cV^*, VD(C_c)), D_c) \in C_{X_2}^{U,Y}$, *where*

$$(VB_c): U \to ((VD(A_c^*))^{(\prime)}, \|\cdot\|')$$

is given by

$$(VB_c)(u)[x] := B_c(u)[V^*x]$$

$u \in U$, $x \in VD(A_c^*)$.

(2) *If* (A_c, B_c, C_c, D_c) *is a state-space realization of the transfer function*

$$G_c(s): \mathrm{RHP} \to L(U, Y),$$

then $(VA_cV^*, VB_c, C_cV^*, D_c)$ *realizes the same transfer function.*

The following definition introduces the standard notation of unitary equivalence of state-space systems. Note that by the previous two propositions, unitarily equivalent systems have the same transfer function.

DEFINITION 5.3. Two systems $(A_c^i, B_c^i, C_c^i, D_c^i) \in C_{X_i}^{U,Y}$, $i = 1, 2$, are called unitarily equivalent, if there exists a unitary operator $V: X_1 \to X_2$ such that

$$(A_c^2, B_c^2, C_c^2, D_c^2) = (VA_c^1V^*, VB_c^1, C_c^1V^*, D_c^1).$$

An equivalent definition applies to admissible discrete-time systems.

We will now show that the transformation $T: D_X^{U,Y} \to C_X^{U,Y}$ and its inverse preserve the unitary equivalence of systems.

PROPOSITION 5.4. *Let* $(A_d^i, B_d^i, C_d^i, D_d^i) \in D_{X_i}^{U,Y}$, $i = 1, 2$. *Let* $(A_c^i, B_c^i, C_c^i, D_c^i) = T((A_d^i, B_d^i, C_d^i, D_d^i))$, $i = 1, 2$, *be the associated continuous-time systems.*

Then, $(A_c^1, B_c^1, C_c^1, D_c^1)$ *and* $(A_c^2, B_c^2, C_c^2, D_c^2)$ *are unitarily equivalent if and only if* $(A_d^1, B_d^1, C_d^1, D_d^1)$ *and* $(A_d^2, B_d^2, C_d^2, D_d^2)$ *are unitarily equivalent.*

Proof. Assume $(A_d^1, B_d^1, C_d^1, D_d^1)$ and $(A_d^2, B_d^2, C_d^2, D_d^2)$ are unitarily equivalent, i.e., there exists a unitary operator $V: X_1 \to X_2$ such that $(A_d^2, B_d^2, C_d^2, D_d^2) = (VA_d^1V^*, VB_d^1, C_d^1V^*, D_d^1)$.

Since $A_d^2 = VA_d^1V^*$ we have

$$A_c^2 = (I + A_d^2)^{-1}(A_d^2 - I) = (I + VA_d^1V^*)^{-1}(VA_d^1V^* - I) = VA_c^1V^*$$

with $D(A_c^2) = VD(A_c^1)$.

Let $u \in U$, $x_2 \in D((A_c^2)^*)$; then

$$B_c^2(u)[x_2] = \sqrt{2}\langle B_d^2(u), (I + (A_d^2)^*)^{-1}x_2\rangle = \sqrt{2}\langle B_d^1(u), (I + (A_d^1)^*)^{-1}V^*x_2\rangle$$

$$= B_c^1(u)[V^*x_2] = [VB_c^1](u)[x_2]$$

and hence $B_c^2 = VB_c^1$.

$C_c^2 = C_c^1V^*$ since for $x \in D(C_c^2)$, we have

$$C_c^2x = \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} \sqrt{2}\, C_d^2(\lambda I + A_d^2)^{-1}x = \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} \sqrt{2}\, C_d^1(\lambda I + A_d^1)^{-1}V^*x = C_c^1V^*x.$$

The fact that $D_c^1 = D_c^2$ follows, since two unitarily equivalent systems have the same transfer function and thus

$$D_c^2 = D_d^2 - \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} C_d^2(\lambda I + A_d^2)^{-1}B_d^2 = D_d^1 - \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} C_d^1(\lambda I + A_d^1)^{-1}B_d^1 = D_c^1.$$

The converse follows similarly. □

6. Dual systems. If (A, B, C, D) is a finite-dimensional linear system, then (A^T, C^T, B^T, D^T) is called the dual system of (A, B, C, D). It is well known that properties of a system are closely related to those of its dual systems. To examine the reachability operator of an infinite-dimensional system via the observability operator of its dual system, we will now define what we mean by the dual system of an admissible system.

We first consider discrete-time systems.

DEFINITION 6.1. Let $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$; then the dual system $(\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)$ of (A_d, B_d, C_d, D_d) is given by

$$\tilde{A}_d := A_d^* : X \to X, \qquad \tilde{B}_c := C_d^* : Y \to X,$$
$$\tilde{C}_d := B_d^* : X \to U, \qquad \tilde{D}_d := D_d^* : Y \to U.$$

The following lemma shows that the dual system of an admissible system is admissible and shows how the transfer function of a system is related to the transfer function of its dual system.

LEMMA 6.2. *The dual system $(\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)$ of an admissible discrete-time system (A_d, B_d, C_d, D_d) in $D_X^{U,Y}$ is an admissible system in $D_X^{Y,U}$.*

If the discrete-time transfer function $G(s): \mathbb{C}\backslash\bar{\mathbf{D}} \to \mathscr{L}(U, Y)$ has an admissible realization (A_d, B_d, C_d, D_d), then the dual system $(\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)$ is a realization of the transfer function $\tilde{G}(s): \mathbb{C}\backslash\bar{\mathbf{D}} \to \mathscr{L}(Y, U)$, $s \mapsto \tilde{G}(s) := (G(\bar{s}))^*$, i.e., for all $s \in \mathbb{C}\backslash\bar{\mathbf{D}}$,

$$\tilde{G}(s) = (G(\bar{s}))^* = \tilde{C}_d(sI - \tilde{A}_d)^{-1}\tilde{B}_d + \tilde{D}_d.$$

Proof. We must check (i)-(v) in Definition 2.1. To show (i) note that since $\|A_d^*\| = \|A_d\|$, we have that A_d^* is a contraction. Thus we only have to show that $-1 \notin \sigma_p(A_d^*)$. Assume there exists $x \in X$ such that $A_d^* x = -x$; then

$$0 \leqq \|A_d x + x\|^2 = \|A_d x\|^2 + 2\,\mathrm{Re}\,\langle x, A_d^* x\rangle + \|x\|^2$$
$$= \|A_d x\|^2 - 2\|x\|^2 + \|x\|^2 = \|A_d x\| - \|x\|^2 \leqq 0.$$

Thus $\|A_d x + x\|^2 = 0$ and hence $A_c x = -x$, which is a contraction to $-1 \notin \sigma_p(A_d)$.

The remaining parts of the lemma are straightforward to check. □

Next, we are going to define the dual system of an admissible continuous-time system.

DEFINITION 6.3. Let $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$. Then the dual system $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ of (A_c, B_c, C_c, D_c) is given by

$(\tilde{A}_c, D(\tilde{A}_c)) = (A_c^*, D(A_c^*))$, the adjoint operator of $(A_c, D(A_c))$;

$\tilde{B}_c : Y \to D(A_c)^{(\prime)}$, $y \mapsto \tilde{B}_c(y)[\,\cdot\,] := \langle y, C_c(\,\cdot\,)\rangle$;

$\tilde{C}_c : D(\tilde{C}_c) \to U$, $D(\tilde{C}_c) = D(\tilde{A}_c) + (I - \tilde{A}_c)^{-1}\tilde{B}_c Y$, where $\tilde{C}_0 x_0$ is defined by

$$\langle u\tilde{C}_c x_0\rangle = B_c(u)[x_0], \quad \text{for } x_0 \in D(A_c^*),\ u \in U,$$

and by

$$\langle \tilde{C}_c x_0 u\rangle = \langle y_0, C_c(I - A_c)^{-1}B_c u\rangle, \quad \text{for } x_0 = (I - \tilde{A}_c)^{-1}\tilde{B}_c y_0,\ y_0 \in Y,\ u \in U;$$

$\tilde{D}_c := D_c^* : Y \to U$.

The following lemma is the continuous-time equivalent of Lemma 6.2.

LEMMA 6.4. *The dual system $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ of an admissible continuous-time system (A_c, B_c, C_c, D_c) is admissible.*

If the continuous-time transfer function $G(s): \mathrm{RHP} \to \mathscr{L}(U, Y)$ has an admissible realization (A_c, B_c, C_c, D_c), then the dual system $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ is a realization of the transfer function $\tilde{G}(s) := (G(\bar{s}))^*$, i.e., for all $s \in \mathrm{RHP}$,

$$\tilde{G}(s) = (G(\bar{s}))^* = \tilde{C}_c(sI - \tilde{A}_c)^{-1}\tilde{B}_c + \tilde{D}_c.$$

Proof. We must show (i)-(vi) of Definition 2.4. Of these, (i)-(iv) and (vi) are straightforward.

Before we show (v) of Definition 2.4 we show the last statement of the lemma. Let $u \in U$, $y \in Y$, and consider for $G(s) = C_c(sI - A_c)^{-1}B_c + D_c$,

$$\langle y, G(s)u\rangle = \langle y, (C_c(sI - A_c)^{-1}B_c + D_c)u\rangle.$$

Using the resolvent identity we have that

$$\begin{aligned}&\langle y, G(s)u\rangle\\ &\quad = \langle y, C_c(I - A_c)^{-1}B_c u\rangle + \langle y, D_c u\rangle + (1 - \bar{s})\langle y, C_c(I - A_c)^{-1}(sI - A_c)^{-1}B_c u\rangle\\ &\quad = \langle \tilde{C}_c(I - \tilde{A}_c)^{-1}\tilde{B}_c y, u\rangle + \langle \tilde{D}_c y, u\rangle + (1 - \bar{s})\langle y, C_c(I - A_c)^{-1}(sI - A_c)^{-1}B_c u\rangle.\end{aligned}$$

Note that for $x \in X$ we have

$$\langle y, C_c(I - A_c)^{-1}x\rangle = \tilde{B}_c(y)[(I - A_c)^{-1}x] = \langle (I - A_c^*)^{-1}\tilde{B}_c y, x\rangle.$$

Using this identity, we now obtain that

$$\begin{aligned}\langle y, C_c(I - A_c)^{-1}(sI - A_c)^{-1}B_c u\rangle &= \langle (I - A_c^*)^{-1}\tilde{B}_c y, (sI - A_c)^{-1}B_c u\rangle\\ &= \overline{\langle (sI - A_c)^{-1}B_c u, (I - A_c^*)^{-1}\tilde{B}_c y\rangle}\\ &= \overline{B_c(u)[(\bar{s}I - A_c^*)^{-1}(I - A_c^*)^{-1}\tilde{B}_c y]}\\ &= \overline{\langle u, \tilde{C}_c(\bar{s}I - A_c^*)^{-1}(I - A_c^*)^{-1}\tilde{B}_c y\rangle}\\ &= \langle \tilde{C}_c(\bar{s}I - \tilde{A}_c)^{-1}(I - \tilde{A}_c)^{-1}\tilde{B}_c y, u\rangle.\end{aligned}$$

Summarizing and again applying the resolvent identity, we have

$$\begin{aligned}&\langle y, G(s)u\rangle\\ &\quad = \langle \tilde{C}_c(I - \tilde{A}_c)^{-1}\tilde{B}_c y, u\rangle + \langle \tilde{D}_c y, u\rangle + (1 - \bar{s})\langle \tilde{C}_c(\bar{s}I - \tilde{A}_c)^{-1}(I - \tilde{A}_c)^{-1}\tilde{B}_c y, u\rangle\\ &\quad = \langle (\tilde{C}_c(\bar{s}I - \tilde{A}_c)^{-1}\tilde{B}_c + \tilde{D}_c)y, u\rangle\\ &\quad = \langle (G(s))^* y, u\rangle.\end{aligned}$$

Hence $(G(\bar{s}))^* = \tilde{C}_c(sI - \tilde{A}_c)^{-1}\tilde{B}_c + \tilde{D}_c$ for all $s \in \mathrm{RHP}$. Now (v) of Definition 2.4 follows, since

$$\lim_{\substack{s\in\mathbb{R}\\ s\to\infty}} \tilde{C}_c(sI - \tilde{A}_c)^{-1}\tilde{B}_c = \lim_{\substack{s\in\mathbb{R}\\ s\to\infty}} (C_c(sI - A_c)^{-1}B_c)^* = \left(\lim_{\substack{s\in\mathbb{R}\\ s\to\infty}} C_c(sI - A_c)^{-1}B_c\right)^* = 0. \qquad \square$$

We will now show that the notion of duality of two systems is carried over between discrete- and continuous-time systems by the transformation T.

PROPOSITION 6.5. *Let* $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$ *and define* $(A_c, B_c, C_c, D_c) := T((A_d, B_d, C_d, D_d))$. *Let* $(A_d^1, B_d^1, C_d^1, D_d^1) \in D_X^{U,Y}$ *be another discrete-time system and let* $(A_c^1, B_c^1, C_c^1, D_c^1) := T((A_d^1, B_d^1, C_d^1, D_d^1))$ *be its corresponding admissible continuous-time system. Then,*

$$(A_d^1, B_d^1, C_d^1, D_d^1) \text{ is the dual system of } (A_d, B_d, C_d, D_d)$$

if and only if

$$(A_c^1, B_c^1, C_c^1, D_c^1) \text{ is the dual system of } (A_c, B_c, C_c, D_c).$$

Proof. Assume $(A_d^1, B_d^1, C_d^1, D_d^1)$ is the dual system of (A_d, B_d, C_d, D_d). Let $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ be the dual system of (A_c, B_c, C_c, D_c). Then $A_d^1 = \tilde{A}_d = A_d^*$ implies that

$$A_c^1 = (I + A_d^1)^{-1}(A_d^1 - I) = ((A_d - I)(I + A_d)^{-1})^* = A_c^* = \tilde{A}_c.$$

For a justification of the operations with the adjoints, see Weidmann [18, p. 74]. The identity $B_d^1 = \tilde{B}_d = C_d^*$ implies for $y \in Y$, $x \in D(A_c)$, that

$$\begin{aligned} B_c^1(y)[x] &= \sqrt{2}\langle B_d^1 y, (I + (A_d^1)^*)^{-1} x\rangle \\ &= \sqrt{2}\langle y, C_d(I + A_d)^{-1} x\rangle \\ &= \langle y, C_c x\rangle \\ &= \tilde{B}_c(y)[x] \end{aligned}$$

and hence $B_c^1 = \tilde{B}_c$.

To show that $C_c^1 = \tilde{C}_c$, we must consider two cases.

(i) For $x \in D(A_c^1)$ we have $C_c^1 x = \tilde{C}_c x$, since for $u \in U$,

$$\begin{aligned} \langle u, C_c^1 x\rangle &= \sqrt{2}\langle u, C_d^1(I + A_d^1)^{-1} x\rangle = \sqrt{2}\langle B_d u, (I + A_d^*)^{-1} x\rangle \\ &= B_c(u)[x] = \langle u, \tilde{C}_c x\rangle. \end{aligned}$$

(ii) Note that for $y_0 \in Y$, $x_0 := (I - A_c^1)^{-1} B_c^1 y_0 = (I - \tilde{A}_c)^{-1} \tilde{B}_c y_0$, since for $x \in D(A_c)$, we have

$$\begin{aligned} &(I - A_c^1)^{-1} B_c^1(y_0)[x] \\ &= \langle (I - A_c^1)^{-1} B_c^1(y_0), x\rangle = \frac{1}{\sqrt{2}}\langle B_d^1 y_0, x\rangle = \frac{1}{\sqrt{2}}\langle y_0, C_d x\rangle \\ &= \langle y_0, C_c(I - A_c)^{-1} x\rangle = \tilde{B}_c(y_0)[(I - A_c)^{-1} x] \\ &= (I - \tilde{A}_c)^{-1} \tilde{B}_c(y_0)[x]. \end{aligned}$$

Then for $u \in U$,

$$\begin{aligned} \langle C_c^1 x_0, u\rangle &= \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} \langle C_d^1(\lambda I + A_d^1)^{-1} B_d^1 y_0, u\rangle = \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} \langle y_0, C_d(\lambda I + A_d)^{-1} B_d u\rangle \\ &= \langle y_0, C_c(I - A_c)^{-1} B_c u\rangle = \langle \tilde{C}_c x_1, u\rangle \\ &= \langle \tilde{C}_c x_0, u\rangle \end{aligned}$$

where the last equality follows since $x_1 := (I - \tilde{A}_c)^{-1} \tilde{B}_c y_0 = x_0$ and hence $C_c^1 = \tilde{C}_c$.

Since $D_d^1 = \tilde{D}_d = \tilde{D}_d^*$, we have that

$$D_c^1 = D_d^1 - \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} C_d^1(\lambda I + A_d^1)^{-1} B_d^1 = D_d^* - \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} B_d^*(\lambda I + A_d^*)^{-1} C_d^* = D_c^* = \tilde{D}_c.$$

Hence we have that $(A_c^1, B_c^1, C_c^1, D_c^1)$ is the dual system of (A_c, B_c, C_c, D_c).

To show the converse, assume that $(A_c^1, B_c^1, C_c^1, D_c^1)$ is the dual system of (A_c, B_c, C_c, D_c) and let $(\tilde{A}_c, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)$ be the dual system of (A_d, B_d, C_d, D_d).

Then

$$A_d^1 = (I + A_c^1)(I - A_c^1)^{-1} = (I + A_c^*)(I - A_c^*)^{-1} = A_d^* = \tilde{A}_d$$

where we apply Theorem 4.19 of Weidmann [18] to justify the manipulations with the adjoints.

The fact that $B_d^1 = \tilde{B}_d$ follows from the following identities, where $y \in Y$, $x \in D(A_c)$,

$$\begin{aligned} \langle B_d^1 y, x\rangle &= \sqrt{2}\langle (I - A_c^1)^{-1} B_c^1 y, x\rangle = \sqrt{2}\, B_c^1(y)[(I - (A_c^1)^*)^{-1} x] \\ &= \sqrt{2}\, \tilde{B}_c(y)[(I - A_c)^{-1} x] = \sqrt{2}\langle y, C_c(I - A_c)^{-1} x\rangle = \langle y, C_d x\rangle \\ &= \langle \tilde{B}_d y, x\rangle. \end{aligned}$$

To show that $C_d^1 = \tilde{C}_d = B_d^*$, let $u \in U$, $x \in D(A_c^*)$,

$$\begin{aligned}\langle u, C_d^1 x\rangle &= \sqrt{2}\langle u, C_c^1(I - A_c^1)^{-1}x\rangle = \sqrt{2}\langle u, \tilde{C}_c(I - A_c^*)^{-1}x\rangle \\ &= \sqrt{2}\, B_c(u)[(I - A_c^*)^{-1}x] = \sqrt{2}\langle (I - A_c)^{-1}B_c u, x\rangle = \langle B_d u, x\rangle \\ &= \langle u, \tilde{C}_d x\rangle.\end{aligned}$$

Since also

$$D_d^1 = C_c^1(I - A_c^1)^{-1}B_c^1 + D_c^1 = (C_c(I - A_c)^{-1}B_c + D_c)^* = D_d^* = \tilde{D}_d,$$

we have the result. □

7. Observability and reachability operators. We are now in a position to discuss some of the central objects of this paper. We define the observability operator for admissible systems. The reachability operator of a system is introduced as the dual of the observability operator of its dual systems.

Having defined observability and controllability gramians of admissible systems, we show one of the main theorems of this paper. It states that the observability operator of a discrete-time system is related by a unitary transformation to the observability operator of its corresponding continuous-time system. This result is the main tool in proving that the transformation T maps discrete-time balanced realizations to continuous-time balanced realizations.

We first define the observability and reachability operators for discrete-time systems.

DEFINITION 7.1. Let $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$; then the operator

$$\begin{aligned}\mathcal{O}_d &: D(\mathcal{O}_d) \to l_Y^2 \\ &x \mapsto (C_d A_d^n x)_{n \geq 0}\end{aligned}$$

is called the observability operator of the system (A_d, B_d, C_d, D_d), where

$$D(\mathcal{O}_d) = \{x \in X \mid (C_d A_d^n x)_{n \geq 0} \in l_Y^2\}.$$

If $\mathcal{O}_d$ is bounded and $\ker(\mathcal{O}_d) = \{0\}$, then the system (A_d, B_d, C_d, D_d) is called observable.

Let $(\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)$ be the dual system of (A_d, B_d, C_d, D_d). If the observability operator $\tilde{\mathcal{O}}_d$ of $(\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)$ is bounded (and hence $D(\tilde{\mathcal{O}}_d) = X$), then the adjoint of $\tilde{\mathcal{O}}_d$ is called the reachability operator $\mathcal{R}_d$ of (A_d, B_d, C_d, D_d), i.e.,

$$\mathcal{R}_d := \tilde{\mathcal{O}}_d^*.$$

If $\mathcal{R}_d$ exists and range $(\mathcal{R}_d)$ is dense in X, the system (A_d, B_d, C_d, D_d) is called reachable.

The analogous definitions for continuous-time systems are now given.

DEFINITION 7.2. Let $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$, then the operator

$$\begin{aligned}\mathcal{O}_c &: D(\mathcal{O}_c) \to L_Y^2([0, \infty[) \\ &x \mapsto C_c e^{tA_c} x\end{aligned}$$

is called the observability operator of the system (A_c, B_c, C_c, D_c), where

$$D(\mathcal{O}_c) = \{x \in X \mid C_c e^{tA_c} x \text{ exists for almost all } t \in [0, \infty[,\ C_c e^{tA_c} x \in L_Y^2([0, \infty[)\}.$$

We say that (A_c, B_c, C_c, D_c) has a bounded observability operator if $D(A_c) \subseteq D(\mathcal{O}_c)$ and $\mathcal{O}_c$ extends to a bounded operator on X. This extension will also be denoted by $\mathcal{O}_c$.

If (A_c, B_c, C_c, D_c) has bounded observability operator $\mathcal{O}_c$ such that $\ker(\mathcal{O}_c) = \{0\}$, then the system (A_c, B_c, C_c, D_c) is called observable.

Let $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ be the dual system of (A_c, B_c, C_c, D_c). If the observability operator $\tilde{\mathcal{O}}_c$ of $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ is a bounded operator on X, the adjoint of $\tilde{\mathcal{O}}_c$ is called the reachability operator $\mathcal{R}_c$ of (A_c, B_c, C_c, D_c), i.e.,

$$\mathcal{R}_c := \tilde{\mathcal{O}}_c^*.$$

If $\mathcal{R}_c$ exists and range $(\mathcal{R}_c)$ is dense in X, the system (A_c, B_c, C_c, D_c) is called reachable.

The notion of reachability and observability gramians as defined below is central in the discussion of balanced realizations in the next section.

DEFINITION 7.3. Let $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$ with bounded reachability operator $\mathcal{R}_d$ and bounded observability operator $\mathcal{O}_d$. Then

$$\mathcal{W}_d := \mathcal{R}_d \mathcal{R}_d^* : X \to X, \qquad \mathcal{M}_d := \mathcal{O}_d^* \mathcal{O}_d : X \to X$$

are called the reachability and the observability gramian, respectively, of the system (A_d, B_d, C_d, D_d). The reachability gramian $\mathcal{W}_c$ and the observability gramian $\mathcal{M}_c$ of a continuous-time system with bounded reachability operator $\mathcal{R}_c$ and observability operator $\mathcal{O}_c$ are similarly defined to be

$$\mathcal{W}_c := \mathcal{R}_c \mathcal{R}_c^* : X \to X, \qquad \mathcal{M}_c := \mathcal{O}_c^* \mathcal{O}_c : X \to X.$$

Before stating the main theorems of this section we present a collection of standard results on Laguerre functions and straightforward modifications thereof. For a reference, see, e.g., Abramowitz and Stegun [1].

PROPOSITION 7.4. *There exists a complete set of orthogonal real-valued functions* $(L_n(t))_{n \geq 0} \subseteq L^2([0, \infty[)$ *such that*

(i) $1/(1+z)\, e^{t((z-1)/(z+1))} = \sum_{n=0}^{\infty} L_n(t) z^n$ *for* $|z| < 1$.

(ii) $\int_0^{\infty} L_n(t) L_m(t)\, dt = \frac{1}{2}\delta_{nm}$ *for all* n, m.

(iii) $|L_n(t)| \leqq 1 \;\; t \in [0, \infty[$, *for* $n \geqq 0$.

(iv) $L_n(t) \in L^1([0, \infty[)$ *for* $n \geqq 0$.

(v) *If* Y *is a separable Hilbert space, then the operator*

$$W : l_Y^2 \to L_Y^2([0, \infty[), \qquad (x_n)_{n \geqq 0} \mapsto \sqrt{2} \sum_{n=0}^{\infty} L_n(t) x_n$$

is unitary, with adjoint

$$W^* : L_Y^2([0, \infty[) \to l_Y^2, \qquad f(t) \mapsto \sqrt{2} \left(\int_0^{\infty} f(t) L_n(t)\, dt \right)_{n \geqq 0}.$$

Now we will state and prove the main theorems of this section. They show that the observability operators of discrete-time systems are related to the observability operators of their corresponding continuous-time systems by a unitary transformation of the input spaces and vice versa. We first consider the case where a discrete-time system is given. Here the connection of its observability operator to the observability operator of its corresponding continuous-time system is investigated.

THEOREM 7.5. *Let* $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$ *and let* $(A_c, B_c, C_c, D_c) := T((A_d, B_d, C_d, D_d))$ *be the corresponding admissible continuous-time system. Then,*

(i) *For* $x \in D((I + A_d)^{-1}) \cap D(\mathcal{O}_d)$, *we have* $x \in D(\mathcal{O}_c)$ *and*

$$\mathcal{O}_c x = W \mathcal{O}_d x$$

where $\mathcal{O}_c$ is the observability operator of (A_c, B_c, C_c, D_c), and W is the unitary operator defined in Proposition 7.4.

(ii) *If $\mathcal{O}_d : X \to l^2_Y$ is bounded, then $\mathcal{O}_c$ extends to a bounded operator given by*

$$\mathcal{O}_c = W\mathcal{O}_d.$$

Proof. (i) Let $x \in D((I+A_d)^{-1}) \cap D(\mathcal{O}_d) = D(A_c) \cap D(\mathcal{O}_d)$. Write

$$F : [0, \infty[\to Y, \qquad t \mapsto F(t) = \sum_{n=0}^{\infty} L_n(t) C_d A_d^n x,$$

$$F_r : [0, \infty[\to Y, \qquad t \mapsto F_r(t) = \sum_{n=0}^{\infty} L_n(t) r^n C_d A_d^n x,$$

$$G : [0, \infty[\to Y, \qquad t \mapsto G(t) = C_c e^{tA_c} x,$$

with $0 < r < 1$. The function $G(t)$, $t \in [0, \infty[$, is well-defined since $e^{tA_c} x \in D(A_c)$ for $x \in D(A_c)$ and hence $e^{tA_c} x \in D(C_c)$.

First note that $F(t)$ is well defined and in $L^2_Y([0, \infty])$, because $(C_d A_d^n x)_{n \geqq 0} \in l^2_Y$ and because $(\sqrt{2}\, L_n(t))_{n \geqq 0}$ forms an orthonormal basis in $L^2([0, \infty[)$.

Now we are going to show that

$$\sqrt{2}\, F_r(t) \to G(t) \quad \text{pointwise weakly as } r \to 1 - 0.$$

Using the notation and results of § 3 we have that $\sum_{n=0}^{\infty} L_n(t) r^n A_d^n = \delta_t(rA_d)$, $0 < r < 1$, since $\delta_t(rz) \in \mathcal{A}$. Hence,

$$F_r(t) = \sum_{n=0}^{\infty} L_n(t) r^n C_d A_d^n x = C_d \left(\sum_{n=0}^{\infty} L_n(t) r^n A_d^n \right)$$

$$= C_d \delta_t(rA_d) = C_d (I + rA_d)^{-1} e^{t(I+rA_d)^{-1}(I-rA_d)} x.$$

Weak convergence now follows from Proposition 3.11.

But $F_r(t) \to F(t)$ in $L^2_Y([0, \infty[)$ as $r \to 1 - 0$, since

$$(r^n C_d A_d^n x)_{n \geqq 0} \to (C_d A_d^n x)_{n \geqq 0} \quad \text{in } l^2_Y \text{ as } r \to 1 - 0.$$

We can now show that these two convergence results imply that for all $y \in Y$

$$\langle y, G(t) \rangle_Y = \langle y, \sqrt{2}\, F(t) \rangle_Y$$

almost everywhere for all $t \in [0, \infty[$. For otherwise, there is an $\varepsilon > 0$ and a measurable set $A \subseteq [0, \infty[$ with Lebesgue measure $\lambda(A) = \varepsilon$ such that

$$\langle y, G(t) \rangle_Y - \langle y, \sqrt{2}\, F(t) \rangle_Y > \varepsilon$$

for $t \in A$. Now, clearly there is an r_0 such that for $r \geqq r_0$

$$\lambda\{t \in A : \langle y, \sqrt{2}\, F_r(t) \rangle - \langle y, \sqrt{2}\, F(t) \rangle > \varepsilon/2\} < \varepsilon/2,$$

and by Egoroff's Theorem, there is an r_1 such that for $r \geqq r_1$

$$\lambda\{t \in A : \langle y, \sqrt{2}\, F_r(t) \rangle - \langle y, G(t) \rangle > \varepsilon/2\} < \varepsilon/2.$$

These three statements together form a contradiction.

Now, since Y is separable we have that

$$C_c e^{tA_c} x = \sqrt{2} \sum_{n=0}^{\infty} L_n(t) C_d A_d^n x$$

almost everywhere for $t \in [0, \infty[$. Thus $\mathcal{O}_c(x) = W\mathcal{O}_d(x)$ for $x \in D(A_c)$.

(ii) Since $\mathcal{O}_d$ is bounded, W is unitary, and $D(A_c)$ is dense in X, $\mathcal{O}_c$ extends to a bounded operator on X. □

The corollary to this theorem shows the equivalent result for reachability operators.

COROLLARY 7.6. *Let* $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$ *and let* $(A_c, B_c, C_c, D_c) := T((A_d, B_d, C_d, D_d))$ *be the corresponding admissible continuous-time system.*

Then, if the reachability operator $\mathcal{R}_d$ *of* (A_d, B_d, C_d, D_d) *exists as a bounded operator, the reachability operator* $\mathcal{R}_c$ *of* (A_c, B_c, C_c, D_c) *exists as a bounded operator and is given by*

$$\mathcal{R}_c = \mathcal{R}_d W^*.$$

Proof. Let $(\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)$ be the dual system of (A_d, B_d, C_d, D_d). By definition $\mathcal{R}_d = \tilde{\mathcal{O}}_d^*$, where $\tilde{\mathcal{O}}_d$ is the observability operator of $(\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)$. Now consider $T((\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)) =: (\tilde{A}_c, \tilde{B}_d, \tilde{C}_c, \tilde{D}_c)$. By Proposition 6.5 we know that $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ is the dual system of (A_c, B_c, C_c, D_c). But the reachability operator $\mathcal{R}_c$ of (A_c, B_c, C_c, D_c) is given by $\mathcal{R}_c = \tilde{\mathcal{O}}_c^*$, where $\tilde{\mathcal{O}}_c$ is the observability operator of $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$. By the previous theorem $\tilde{\mathcal{O}}_c = W\tilde{\mathcal{O}}_d$ and hence $\mathcal{R}_c = \tilde{\mathcal{O}}_c^* = \tilde{\mathcal{O}}_d^* W^* = \mathcal{R}_d W^*$. □

We now show that if a continuous-time system has a bounded observability operator $\mathcal{O}_c$ then the observability operator of its corresponding discrete time system is given by a unitary transformation of $\mathcal{O}_c$.

THEOREM 7.7. *Let* $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$ *and let* $(A_d, B_d, C_d, D_d) = T^{-1}((A_c, B_c, C_c, D_c))$ *be the corresponding admissible discrete time system. Then,*

(i) *For* $x \in D(A_c) \cap D(\mathcal{O}_c)$, *we have* $x \in D(\mathcal{O}_d)$ *and*

$$\mathcal{O}_d x = W^* \mathcal{O}_c x$$

where $\mathcal{O}_d$ *is the observability operator of* (A_d, B_d, C_d, D_d) *and* W^* *is the unitary operator defined in Proposition* 7.4.

(ii) *If* $\mathcal{O}_c$ *is bounded, then* $\mathcal{O}_d$ *extends to a bounded operator on* X *given by*

$$\mathcal{O}_d = W^* \mathcal{O}_c.$$

Proof. (i) Let $x \in D(A_c) \cap D(\mathcal{O}_c)$; then we know that $G(t) := C_c e^{tA_c} x$ exists for all $t \in [0, \infty]$, since $e^{tA_c} x \in D(A_c) \subseteq D(C_c)$, $t \in [0, \infty[$. By assumption $G(t) \in L_Y^2([0, \infty[)$.

Corollary 4.4 implies that

$$G(t) = \sqrt{2}\, C_d (I + A_d)^{-1} e^{t(I+A_d)^{-1}(A_d - I)} x.$$

For $0 < r < 1$, let $G_r(t) = \sqrt{2}\, C_d (I + rA_d)^{-1} e^{t(I+rA_d)^{-1}(rA_d - I)} x$. Since C_d is bounded, we have by Proposition 3.11 that for all $t \in [0, \infty[$,

$$\lim_{r \to 1-0} G_r(t) = G(t) \quad \text{weakly.}$$

Since C_d is bounded and $\delta_t(rz) \in \mathcal{A}$, where δ_t is as defined in § 3, we have that

$$\begin{aligned} G_r(t) &= \sqrt{2}\, C_d (I + rA_d)^{-1} e^{t(I+rA_d)^{-1}(rA_d - I)} x \\ &= \sqrt{2}\, C_d \delta_t(rA_d) x \\ &= \sqrt{2}\, C_d \left(\sum_{n=0}^{\infty} L_n(t) r^n A_d^n x \right) \\ &= \sqrt{2} \sum_{n=0}^{\infty} L_n(t) r^n C_d A_d^n x \in L_Y^2([0, \infty]). \end{aligned}$$

We will now show that there exists $M>0$ such that

$$\|G_r(t)\| \leqq M < \infty \quad \text{for all } 0<r<1, \quad t \in [0, \infty[.$$

Let $t \in [0, \infty[$; then

$$\begin{aligned}\sup_{0<r<1} \|G_r(t)\| &= \sup_{0<r<1} \|\sqrt{2}\, C_d (I+rA_d)^{-1}\, e^{t(I+rA_d)^{-1}(rA_d-I)}x\| \\ &\leqq \sqrt{2}\, \|C_d\| \sup_{0<r<1} \|(I+rA_d)^{-1}\, e^{t(I+rA_d)^{-1}(rA_d-I)}x\| \\ &\leqq M < \infty\end{aligned}$$

where the second to last line follows from Proposition 3.11 noting that $D(A_c) = D((I+A_d)^{-1})$.

Thus for $y \in Y$, we have that

$$|\langle G_r(t), y\rangle \sqrt{2}\, L_n(t)| \leqq \sqrt{2}\, M \|y\|\, |L_n(t)|, \qquad t \in [0, \infty[.$$

Since $L_n(t) \in L^1([0, \infty[)$, we can therefore apply the Dominated Convergence Theorem:

$$\lim_{r \to 1-0} \int_0^\infty \langle G_r(t), y\rangle \sqrt{2}\, L_n(t)\, dt = \int_0^\infty \langle G(t), y\rangle \sqrt{2}\, L_n(t)\, dt.$$

But

$$\begin{aligned}\int_0^\infty \langle G_r(t), y\rangle \sqrt{2}\, L_n(t)\, dt &= \int_0^\infty \left\langle \sum_{i=0}^\infty \sqrt{2}\, L_i(t) r^i C_d A_d^i x, y \right\rangle \sqrt{2}\, L_n(t)\, dt \\ &= 2 \sum_{i=0}^\infty \left(r^i \langle C_d A_d^i x, y\rangle \int_0^\infty L_i(t) L_n(t)\, dt \right) \\ &= r^n \langle C_d A_d^n x, y\rangle.\end{aligned}$$

Thus

$$\begin{aligned}\int_0^\infty \langle G(t), y\rangle \sqrt{2}\, L_n(t)\, dt &= \lim_{r \to 1-0} \int_0^\infty \langle G_r(t), y\rangle \sqrt{2}\, L_n(t)\, dt \\ &= \langle C_d A_d^n x, y\rangle.\end{aligned}$$

Since $G(t) \in L_Y^2([0, \infty[)$, we have an expansion

$$G(t) = \sum_{n=0}^\infty G_n \sqrt{2}\, L_n(t), \qquad G_n \in Y.$$

Thus $\langle G_n, y\rangle = \langle C_d A_d^n x, y\rangle$ for all $y \in Y$ and hence $G_n = C_d A_d^n x$ for $n \geqq 0$. This implies that

$$\mathcal{O}_c(x) = C_c\, e^{tA_c} x = \sqrt{2} \sum_{n=0}^\infty L_n(t) C_d A_d^n x = W \mathcal{O}_d(x)$$

and hence $\mathcal{O}_d(x) = W^* \mathcal{O}_c(x)$.

(ii) This is a straightforward consequence of (i). □

In the following corollary the corresponding result is established for the reachability operators.

COROLLARY 7.8. *Let* $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$ *and let* $(A_d, B_d, C_d, D_d) := T^{-1}((A_c, B_c, C_c, D_c))$ *be the corresponding admissible discrete-time system. Then, if the reachability operator* $\mathcal{R}_c$ *of* (A_c, B_c, C_c, D_c) *exists as bounded operator, the reachability operator* $\mathcal{R}_d$ *of* (A_d, B_d, C_d, D_d) *exists as a bounded operator and is given by*

$$\mathcal{R}_d = \mathcal{R}_c W.$$

Proof. Let $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ be the dual system of (A_c, B_c, C_c, D_c). By definition, $\mathcal{R}_c = \tilde{\mathcal{O}}_c^*$, where $\tilde{\mathcal{O}}_c$ is the observability operator of $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$. Now consider $(\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d) = T^{-1}((\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c))$. By Proposition 6.5 we know that $(\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)$ is the dual system of (A_d, B_d, C_d, D_d). But the reachability operator $\mathcal{R}_d$ of (A_d, B_d, C_d, D_d) is given by $\mathcal{R}_d = \tilde{\mathcal{O}}_d^*$, where $\tilde{\mathcal{O}}_d$ is the observability operator of $(\tilde{A}_d, \tilde{B}_d, \tilde{C}_d, \tilde{D}_d)$. By the previous theorem $\tilde{\mathcal{O}}_d = W^* \tilde{\mathcal{O}}_c$. Hence $\mathcal{R}_d = \tilde{\mathcal{O}}_d^* = \tilde{\mathcal{O}}_c^* W = \mathcal{R}_c W$. □

The following corollary to the previous two theorems shows that the properties of observability and reachability as well as the observability and reachability gramians are preserved by the transformation T.

COROLLARY 7.9. *Let* $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$ *and* $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$ *such that*

$$(A_c, B_c, C_c, D_c) = T((A_d, B_d, C_d, D_d)).$$

Then,

(1) (A_c, B_c, C_c, D_c) *is observable* (*reachable*) *if and only if* (A_d, B_d, C_d, D_d) *is observable* (*reachable*).

(2) *If the reachability gramians* $\mathcal{W}_c$, $\mathcal{W}_d$ (*observability gramians* $\mathcal{M}_c$, $\mathcal{M}_d$) *of* (A_d, B_d, C_d, D_d) *and* (A_c, B_c, C_c, D_c) *are defined, then*

$$\mathcal{W}_c = \mathcal{W}_d \qquad (\mathcal{M}_c = \mathcal{M}_d).$$

8. Balanced realizations. We will now apply the results on infinite-dimensional state-space systems of the previous sections to tackle the problem that motivated this paper, namely, that of the existence of balanced realizations for continuous-time systems. Our results will allow us to deal with a wider range of transfer functions than previous results; for example, we can handle any transfer function that is bounded in the RHP, and with a limit at infinity along the real axis. This allows us to consider nonstrictly proper delay systems with transfer functions such as $G(s)\, e^{-sT}$, where $G(s)$ is a matrix-valued stable rational transfer function. Previous results were unable to deal with, for example, the pure delay system e^{-sT} because the limits $\lim_{w \to +\infty} e^{-iwt}$ and $\lim_{w \to -\infty} e^{-iwt}$ do not exist and, therefore, the corresponding Hankel operator is not compact. Another example of a function we will be able to deal with is $G(s) = \log(1 + 1/s)$. This function is unusual in that it has a singularity at 0.

The approach taken is to carry over the discrete-time results by Young using the transformation $T: D_X^{U,Y} \to C_X^{U,Y}$. Thus we will first review Young's results before we turn to proving the continuous-time analogue of his discrete-time realization theorem.

The following definition recalls the notion of a balanced system as defined by Moore [10] and the notion of a parbalanced system as introduced by Young [20].

DEFINITION 8.1. Let $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$ $((A_c, B_c, C_c, D_c) \in C_X^{U,Y})$ be such that the observability gramian $\mathcal{M}_d$ $(\mathcal{M}_c)$ and reachability gramian $\mathcal{W}_d$ $(\mathcal{W}_c)$ exist. Then the system is

(i) Parbalanced, if $\mathcal{M}_d = \mathcal{W}_d$ $(\mathcal{M}_c = \mathcal{W}_c)$;

(ii) Balanced, if it is parbalanced and moreover the gramians are diagonal.

Before we state any results, we introduce some notation. Let $H: \mathbf{D} \to \mathcal{L}(U, Y)$ be analytic. We say that $H \in P_+ L^\infty(\mathbf{D}, \mathcal{L}(U, Y))$ if there exists an analytic function $F: \mathbf{D} \to \mathcal{L}(U, Y)$ such that $H + \bar{F}$ is essentially bounded, where $\bar{F}(z) = F(z^{-1})$. Furthermore, if F can be chosen so that $H + \bar{F} \in C(\mathbf{D}, \mathcal{K}(U, Y))$, where $C(\mathbf{D}, \mathcal{K}(U, Y))$ is the set of norm continuous functions on $\partial\mathbf{D}$ with values in the set of compact operators from U to Y, then H is said to be in $P_+ C(\mathbf{D}, \mathcal{K}(U, Y))$.

Similarly, if $H: \mathrm{RHP} \to \mathscr{L}(U, Y)$ is analytic, we say that $H \in P_+L^\infty(\mathrm{RHP}, \mathscr{L}(U, Y))$ ($P_+C(\mathrm{RHP}, \mathscr{K}(U, Y))$) if there is an analytic function $F: \mathrm{RHP} \to \mathscr{L}(U, Y)(\mathscr{K}(U, Y))$ such that $H + \tilde{F}$ is essentially bounded (extends to a norm continuous function on the imaginary axis such that $\lim_{w \in \mathbb{R}, w \to \infty} (H + \tilde{F})(iw) = \lim_{w \in \mathbb{R}, w \to \infty} (H + \tilde{F})(-iw)$), where $\tilde{F}(s) = F(-s)$.

Remark 8.2. If $H \in P_+L^\infty(\mathbf{D}, \mathscr{L}(U, Y))$, then the Hankel operator with symbol H is bounded by an operator-valued version of Nehari's Theorem, whereas by Hartmann's Theorem it is compact if $H \in P_+C(\mathbf{D}, \mathscr{L}(U, Y))$. Note that if U and Y are finite-dimensional, $H \in P_+L^\infty(\mathbf{D}, \mathscr{L}(U, Y))$ ($P_+L^\infty(\mathrm{RHP}, \mathscr{L}(U, Y))$) if and only if H is in $BMOA(\partial\mathbf{D})$ ($BMOA(i\mathbb{R})$) and $H \in P_+C(\mathbf{D}, \mathscr{K}(U, Y))$ ($P_+C(\mathrm{RHP}, \mathscr{K}(U, Y))$) if and only if H is in $VMOA(\partial\mathbf{D})$ ($VMOA(i\mathbb{R})$) (for references, see [15]).

The following theorem by Young [20], gives criteria for a (par-) balanced realization to exist of a discrete-time transfer function.

THEOREM 8.3. *Let $G_d(z): \mathbb{C} \setminus \bar{\mathbf{D}} \to \mathscr{L}(U, Y)$ be analytic with $G_d(\infty) = D_d \in \mathscr{L}(U, Y)$, and write*

$$g(z) := \frac{1}{z}\left(G_d\left(\frac{1}{z}\right) - D_d\right), \qquad z \in \mathbf{D}.$$

(i) *If $g \in P_+L^\infty(\mathbf{D}, \mathscr{L}(U, Y))$, then there exists a separable Hilbert space X and a discrete-time state-space realization (A_d, B_d, C_d, D_d) of $G_d(z)$ with state space X, such that*

$$A_d \in \mathscr{L}(X) \quad \text{is a contraction,}$$

$$B_d \in \mathscr{L}(U, X), \qquad C_d \in \mathscr{L}(X, Y),$$

and (A_d, B_d, C_d, D_d) is reachable and observable with bounded reachability and observability operators, such that (A_d, B_d, C_d, D_d) is parbalanced, i.e., $\mathscr{M}_d = \mathscr{W}_d$. The gramians $\mathscr{M}_d$, $\mathscr{W}_d$ satisfy the Lyapunov equations

$$A_d \mathscr{W}_d A_d^* - \mathscr{W}_d = -B_d^* B_d, \qquad A_d^* \mathscr{M}_d A_d - \mathscr{M}_d = -C_d C_d^*.$$

If $(\bar{A}_d, \bar{B}_d, \bar{C}_d, \bar{D}_d)$ is another parbalanced realization of $G_d(z)$ with state space $\bar{X}$, then (A_d, B_d, C_d, D_d) and $(\bar{A}_d, \bar{B}_d, \bar{C}_d, \bar{D}_d)$ are unitarily equivalent.

(ii) *If, moreover, $g \in P_+C(\mathbf{D}, \mathscr{K}(U, Y))$, there is a basis in X with respect to which (A_d, B_d, C_d, D_d) is balanced.*

To show that for a transfer function G_d such that $\lim_{\lambda < -1, \lambda \to -1} G_d(\lambda) \in \mathscr{L}(U, Y)$, the realization given in the previous theorem is, in fact, admissible, we need to show that -1 is not an eigenvalue of A_d.

LEMMA 8.4. *Let (A_d, B_d, C_d, D_d) be a parbalanced realization of a discrete-time transfer function as given in Theorem* 8.3; *then $-1 \notin \sigma_p(A_d)$.*

Proof. Let $\mathscr{M}_d$ be the observability gramian of (A_d, B_d, C_d, D_d); then

$$A_d^* \mathscr{M}_d A_d - \mathscr{M}_d = -C_d^* C_d.$$

Assume $-1 \in \sigma_p(A_d)$ with eigenvector $x \neq 0$, then

$$\langle x, A_d^* \mathscr{M}_d A_d x \rangle - \langle x, \mathscr{M}_d x \rangle = -\langle x, C_d^* C_d x \rangle$$

and hence

$$\langle A_d x, \mathscr{M}_d A_d x \rangle - \langle x, \mathscr{M}_d x \rangle = 0 = -\| C_d x \|^2,$$

which implies that $C_d x = 0$. Hence for all $n \geqq 0$, $\mathscr{O}_d x = (C_d A_d^n x)_{n \geqq 0} = (-1)^n C_d x = 0$, which is a contradiction to the observability of (A_d, B_d, C_d, D_d). □

We can now apply our results on the transformation T to obtain realization results for continuous-time transfer functions.

THEOREM 8.5. *Let $G_c: \mathrm{RHP} \to \mathscr{L}(U, Y)$ be a continuous-time transfer function that is analytic and such that $\lim_{s\in\mathbb{R}, s\to\infty} G_c(s) \in \mathscr{L}(U, Y)$ exists.*

(i) *If $G_c \in P_+L^\infty(\mathrm{RHP}, \mathscr{L}(U, Y))$, then there exists a separable Hilbert space X and a parbalanced admissible continuous-time state-space realization (A_c, B_c, C_c, D_c) of G_c with state space X. This system is reachable, observable, and has bounded reachability and observability operators.*

If $(\bar{A}_c, \bar{B}_c, \bar{C}_c, \bar{D}_c)$ is another parbalanced realization of $G_c(s)$, then (A_c, B_c, C_c, D_c) and $(\bar{A}_c, \bar{B}_c, \bar{C}_c, \bar{D}_c)$ are unitarily equivalent.

(ii) *If, moreover, $G_c \in P_+C(\mathrm{RHP}, \mathscr{K}(U, Y))$, then there is a basis in X with respect to which (A_c, B_c, C_c, D_c) is balanced.*

Proof. Let $G_d: \mathbb{C}\backslash\bar{\mathbf{D}} \to \mathscr{L}(\mathbf{U}, \mathbf{Y})$ be the associated discrete-time transfer function $G_d(z) = G_c((z-1)/(z+1))$, and write

$$g(z) = \frac{1}{z}\left(G_d\left(\frac{1}{z}\right) - G_c(1)\right), \qquad z \in \mathbf{D}.$$

Then it is easy to see that $G_c \in P_+L^\infty(\mathrm{RHP}, \mathscr{L}(U, Y))(P_+C(\mathrm{RHP}, \mathscr{K}(U, Y)))$ if and only if $g \in P_+L^\infty(\mathbf{D}, \mathscr{L}(U, Y))(P_+C(\mathbf{D}, \mathscr{K}(U, Y)))$.

Hence $G_d(z)$ has a parbalanced realization (A_d, B_d, C_d, D_d) that is admissible since A_d is a contraction, such that $-1 \notin \sigma_p(A_d)$ by Lemma 8.4 and since

$$\lim_{\substack{\lambda<1\\ \lambda\to 1}} C_d(\lambda I + A_d)^{-1}B_d = -\lim_{\substack{\lambda<-1\\ \lambda\to -1}} G_d(\lambda) + D_d = -\lim_{\substack{s\in\mathbb{R}\\ s\to\infty}} G_c(s) + D_d \in \mathscr{L}(U, Y)$$

exists. Then $(A_c, B_c, C_c, D_c) := T((A_d, B_d, C_d, D_d)) \in C_X^{U,Y}$ is a state-space realization of G_c (Theorem 4.1) that is observable and reachable, such that $\mathscr{W}_c = \mathscr{W}_d$ and $\mathscr{M}_c = \mathscr{M}_d$ (Corollary 7.9).

The statement on the uniqueness of the realization follows from Proposition 5.4. □

The following corollary discusses special cases of transfer functions and gives simple criteria for the existence of a parbalanced or balanced realization of a continuous-time transfer function.

COROLLARY 8.6. *Let $G_c(s): \mathrm{RHP} \to \mathscr{L}(U, Y)$ be a continuous-time transfer function, such that $\lim_{s\in\mathbb{R}, s\to\infty} G_c(s) \in \mathscr{L}(U, Y)$ exists and $G_c(s)$ is analytic in* RHP.

(i) *If $G_c(s)$ is bounded in the* RHP, *i.e.*, $\sup_{s\in\mathrm{RHP}} \|G_c(s)\| < \infty$, *then $G_c(s)$ has a parbalanced realization.*

(ii) *If, in particular, $G_c(s): \mathrm{RHP} \mapsto \mathscr{K}(U, Y)$, such that G_c is bounded in the* RHP *and $G_c(s)$ is norm continuous on the imaginary axis including at the points $+\infty$ and $-\infty$, i.e., $w \mapsto G_c(iw)$, $w \in \mathbb{R}$, is norm continuous and $\lim_{w\to-\infty} G_c(iw) = \lim_{w\to+\infty} G_c(iw)$, then $G_c(s)$ has a balanced realization.*

As examples to the previous realization results we can consider delay systems. It follows immediately from the previous corollary that the transfer function e^{-sT} of a pure delay with time constant $T > 0$, has a parbalanced realization. Note that the limits $\lim_{w\to+\infty} e^{-iwT}$ and $\lim_{w\to-\infty} e^{-iwT}$ do not exist. If $G(s)$ is a matrix-valued strictly proper stable rational transfer function, then the transfer function $G(s)\, e^{-sT}$ of the delayed system has a balanced realization.

Another example of a function that has a parbalanced continuous-time state-space realization is the function $G_c(s) = \log(1 + 1/s)$, which is well known to be in $BMOA(i\mathbb{R})$. We note that G_c has a singularity at 0.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1964.

[2] R. F. CURTAIN AND K. GLOVER, *Balanced Realizations for Infinite Dimensional Systems*, Operator Theory, Advances and Applications No. 19, Birkhäuser Verlag, Boston, 1986.

[3] B. A. FRANCIS, *A course in H^∞ Control Theory*, Lecture Notes in Control and Information Science, Springer-Verlag, Berlin, New York, 1987.

[4] P. A. FUHRMANN, *Linear Systems and Operators in Hilbert Space*, McGraw-Hill, New York, 1981.

[5] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.

[6] K. GLOVER, R. F. CURTAIN, AND J. R. PARTINGTON, *Realisation and Approximation of linear infinite dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863-898.

[7] S. J. HEGNER, *Linear decomposable systems in continuous time*, SIAM J. Math. Anal., 12 (1981), pp. 243-273.

[8] D. G. HEDBERG, *Operator models of infinite dimensional systems*, Ph.D. Thesis, Department of Engineering, University of California, Los Angeles, CA, 1977.

[9] J. W. HELTON, *Systems with infinite-dimensional state space. The Hilbert space approach*, Proc. IEEE, 64 (1976), pp. 145-160.

[10] B. C. MOORE, *Principal component analysis in linear systems: controllability, observability and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17-32.

[11] R. OBER, *Balanced realizations: canonical form, parametrization, model reduction*, Internat. J. Control, 46 (1987), pp. 643-670.

[12] ———, *A parametrization approach to infinite dimensional balanced systems and their approximation*, IMA J. Math. Control Inform., 4 (1987), pp. 263-279.

[13] ———, *The parametrization of linear systems using balanced realizations: relaxation systems*, Linear Circuits, Systems and Signal Processing, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North Holland, Amsterdam, 1988, pp. 313-320.

[14] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.

[15] S. C. POWER, *Hankel operators on Hilbert space*, Bull. London Math. Soc., 12 (1980), pp. 422-442.

[16] D. SALAMON, *Realization theory in Hilbert space*, Mathematical Systems Theory, 21 (1989), pp. 147-164.

[17] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North Holland, Amsterdam, 1970.

[18] J. WEIDMANN, *Linear Operators in Hilbert Spaces*, Springer-Verlag, Berlin, New York, 1980.

[19a] Y. YAMAMOTO, *Realization theory of infinite-dimensional systems, Part* I, Math. Systems Theory, 15 (1981), pp. 55-77.

[19b] ———, *Realization theory of infinite-dimensional systems, Part* II, Math. Systems Theory, 15 (1982), pp. 169-190.

[20] N. YOUNG, *Balanced Realizations in Infinite Dimensions*, Operator Theory, Advances and Applications No. 19, Birkhäuser Verlag, Boston, 1986.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 466-477, March 1990

012

UNIFORM STABILIZATION OF THE WAVE EQUATION BY NONLINEAR BOUNDARY FEEDBACK*

ENRIKE ZUAZUA†

Abstract. The question of uniformly stabilizing the solution of the wave equation $y''-\Delta y=0$ in $\Omega\times(0,\infty)$ (Ω is a bounded domain of $\mathbf{R}^n$) by means of a nonlinear feedback law of the following form is studied: $\partial y/\partial v=-k(x)g(y')$ on $\Gamma_0\times(0,\infty)$, $y=0$ on $\Gamma_1\times(0,\infty)$, (Γ_0,Γ_1) being a suitable partition of the boundary of Ω and g a continuous nondecreasing function such that $g(0)=0$. We choose $k(x)\in L^\infty(\Gamma_0)$, $k(x)\geqq 0$ such that $k(x)$ vanishes linearly at the interface points $x\in\bar{\Gamma}_0\cap\bar{\Gamma}_1$. Then, if $g(s)$ behaves like $|s|^{p-1}s$ as $|s|\to 0$ with $p>1$ and linearly as $|s|\to\infty$, it is proved that the energy of every solution decays like $t^{-2/(p-1)}$ as $t\to\infty$. In the case where $p=1$ the exponential decay rate is proved.

Key words. wave equation, boundary damping, nonlinear feedback law, uniform stabilization, decay rate estimate

AMS(MOS) subject classifications. 35B40, 93D15

1. Introduction. Let Ω be a bounded, open, connected set in $\mathbf{R}^n$ ($n\geqq 1$) having a boundary $\Gamma=\partial\Omega$ of class C^2. Given a point $x^0\in\mathbf{R}^n$, let be $m(x)=x-x^0$ and consider the following partition of the boundary:

$$\Gamma(x^0)=\{x\in\Gamma: m(x)\cdot\nu(x)>0\}, \tag{1.1}$$

$$\Gamma_*(x^0)=\{x\in\Gamma: m(x)\cdot\nu(x)\leqq 0\}=\Gamma/\Gamma(x^0) \tag{1.2}$$

where $\nu(x)$ is the unit normal vector to Γ at $x\in\Gamma$ pointing toward the exterior of Ω and " $\cdot$ " denotes the scalar product in $\mathbf{R}^n$. We will denote by $\partial/\partial\nu$ the normal derivative in the direction ν and by $'=\partial^*/\partial t$ the time-derivative.

Let us assume that $\operatorname{int}\Gamma_*(x^0)\neq\varnothing$ and consider the following wave equation:

$$y''-\Delta y=0 \qquad \text{in } \Omega\times(0,\infty), \tag{1.3}$$

$$\frac{\partial y}{\partial\nu}=-\{m(x)\cdot\nu(x)\}g(y') \qquad \text{on } \Gamma(x^0)\times(0,\infty), \tag{1.4}$$

$$y=0 \qquad \text{on } \Gamma_*(x^0)\times(0,\infty), \tag{1.5}$$

$$y(0)=y^0\in V, \qquad y'(0)=y^1\in L^2(\Omega) \tag{1.6}$$

where

$$V=\{\phi\in H^1(\Omega): \phi=0 \text{ on } \Gamma_*(x^0)\} \tag{1.7}$$

and $g\in C(\mathbf{R})$ is a nondecreasing function such that $g(0)=0$ and $g(s)s>0$ for all $s\in\mathbf{R}-\{0\}$.

We define the energy of a solution $y=y(x,t)$ of (1.3)-(1.6) as follows:

$$E(t)=\frac{1}{2}\int_\Omega\{|\nabla y(x,t)|^2+|y'(x,t)|^2\}\,dx \quad \forall t\geqq 0. \tag{1.8}$$

* Received by the editors August 3, 1988; accepted for publication (in revised form) March 24, 1989. Part of this work was done when the author was visiting the Department of Mathematics, Georgetown University, Washington, DC. This work was partially supported by grant G. V. 127.310-1/87 of the "Eusko Jaurlaritza" (Basque Government).

† Departamento de Matemáticas, Universidad del Pais Vasco, Apartado 644, 48080 Bilbao, Spain.

Let us calculate formally (all this will be justified below) the derivative of the energy. By using Green's formula we easily obtain

$$E'(t)=\frac{dE(t)}{dt}=-\int_{\Gamma(x^0)}\{m(x)\cdot\nu(x)\}g(y'(x,t))y'(x,t)\,d\Gamma \tag{1.9}$$

where $d\Gamma$ is the surface measure associated to the boundary Γ. Taking into account the fact that $g(s)s\geqq 0$, for all $s\in\mathbf{R}$ we deduce from (1.9) that the energy $E(t)$ is nonincreasing.

Applying La Salle's invariance principle it is easy to see that the energy goes to zero as t goes to infinity. Indeed, in addition to (1.9) we have

$$\begin{aligned}&\frac{d}{dt}\int_\Omega\{|\nabla(y_1-y_2)(x,t)|^2+|(y_1'-y_2')(x,t)|^2\}\,dx\\ &=-2\int_{\Gamma(x^0)}\{m(x)\cdot\nu(x)\}(g(y_1'(x,t))-g(y_2'(x,t)))(y_1'-y_2')(x,t)\,d\Gamma \qquad (1.10)\\ &\leqq 0\end{aligned}$$

for any two finite energy solutions y_1, y_2 of (1.3)-(1.6).

It is enough to establish the decay for a dense subset of initial data, for instance,

$$W=\{(y^0,y^1)\in V\times V:\Delta y^0\in L^2(\Omega),\frac{\partial y^0}{\partial\nu}=-(m\cdot\nu)g(y^1)\text{ on }\Gamma(x^0)\}.$$

Let us define the higher-order energy

$$F(t)=\frac{1}{2}\int_\Omega(|\Delta y(x,t)|^2+|\nabla y'(x,t)|^2)\,dx. \tag{1.11}$$

We have

$$\frac{dF(t)}{dt}=-\int_{\Gamma(x^0)}\{m\cdot\nu\}g'(y')|y''|^2\,d\Gamma\leqq 0 \tag{1.12}$$

and thus

$$F(t)\leqq F(0)<\infty\quad\forall t\geqq 0. \tag{1.13}$$

Therefore the trajectory $\{y(t),y'(t)\}$ corresponding to the initial data in W is relatively compact in $V\times H$. Applying La Salle's invariance principle we introduce the ω-limit set $\omega\{y^0,y^1\}$ with respect to the strong topology of $V\times H$, and the problem reduces to prove that

$$\omega\{y^0,y^1\}=\{0,0\}. \tag{1.14}$$

From (1.9) we easily deduce that $\omega\{y^0,y^1\}$ is contained in the set of initial data leading to solutions with constant energy, i.e., such that

$$\begin{aligned}&y''-\Delta y=0 && \text{in }\Omega\times(0,\infty),\\ &\frac{\partial y}{\partial\nu}=y'=0 && \text{on }\Gamma(x^0)\times(0,\infty), \qquad (1.15)\\ &y=0 && \text{on }\Gamma_*(x^0)\times(0,\infty).\end{aligned}$$

But Holmgren's Uniqueness Theorem assures that the only function satisfying (1.15) is the trivial one $y=0$, and thus (1.14) holds.

To obtain the compactness of trajectories we have used in (1.12) the fact that g is locally Lipschitz with $g' \geqq 0$ almost everywhere. However, the compactness can be achieved for all nondecreasing continuous function g such that $g(s)s > 0$, for all $s \in \mathbf{R} - \{0\}$ by obtaining uniform energy estimates for the family of functions

$$z_h(x, t) = \frac{y(x, t+h) - y(x, t)}{h}$$

as $h \to 0$.

We note that more general stability results have been recently proven by Chen and Wang in [6] and Lasiecka in [17], [18] in the case where g is a multivalued maximal monotone function.

The aim of this paper is to estimate the rate of decay of the energy and, more precisely, to establish some relations between the rate of the decay and the behavior of the nonlinearity g.

The linear case $g(s) = as$, with $a \in \mathbf{R}$, is well known by now. The first stabilization result is due to Chen [3] (see also [4], [5]). He proved the exponential decay of the energy in this linear case with boundary conditions

$$\frac{\partial y}{\partial \nu} = -k(x)y' \quad \text{on } \Gamma(x^0) \times (0, \infty) \tag{1.16}$$

with $k \in L^\infty(\Gamma(x^0))$, $k \geqq k_0 > 0$ instead of (1.4) but only under the geometrical restriction

$$(x - x^0) \cdot \nu(x) \geqq \gamma > 0 \quad \text{on } \Gamma(x^0). \tag{1.17}$$

Chen's result was later generalized by Lagnese in [13] by considering a more general class of multipliers (other than the radial one $x - x^0$) but always under a geometrical restriction of type (1.17).

An important observation is that condition (1.17) forces

$$\overline{\Gamma(x^0)} \cap \overline{\Gamma_*(x^0)} = \varnothing. \tag{1.18}$$

Thus if $\Gamma(x^0) \neq \varnothing$, the above results cannot apply to regions Ω having a smooth connected boundary. However, in a recent paper by Komornik and Zuazua [11], [12] the geometrical restriction (1.17) has been avoided and the exponential decay has been proved for dimensions $n \leqq 3$ (for some technical reasons that we will discuss in § 2) by taking in (1.16) the weight

$$k(x) = (x - x^0) \cdot \nu(x). \tag{1.19}$$

Subsequently, Lagnese in [15] generalized this result to a larger class of multipliers but always considered weights of type (1.19) in the boundary condition (1.16).

As mentioned above, the first contribution in [11] and [12] was that, for the first time, the geometrical restriction (1.17) was avoided and the second was the method by itself. Indeed, in all preceding works the exponential decay of the energy was obtained from estimates on $\int_0^\infty E(t)\, dt$ by employing a result of Datko [7]. However, in [11] and [12] a different point of view was taken and the exponential decay was obtained by constructing perturbed energy functionals for which differential inequalities leading to the exponential decay were obtained. The advantage of this approach is that it can be easily extended to treat some semilinear problems as in [12]. We note also that Lagnese proved in a recent work [16], inspired by these kinds of methods, stabilization results for von Karman's plate models.

The aim of the present paper is to show how this method can be adapted to obtain decay estimates for wave equations with nonlinear dissipative boundary conditions.

We will consider here functions g satisfying the following conditions:

$$|g(s)| \leqq C_1|s| \quad \forall s \in \mathbf{R}, \tag{1.20}$$

$$|g(s)| \geqq C_2 \inf\{|s|, |s|^p\} \quad \forall s \in \mathbf{R} \tag{1.21}$$

for some C_1, $C_2 > 0$, and $p \geqq 1$.

Under these assumptions, we will prove the following:

(a) When $p = 1$ the energy of every solution decays exponentially;

(b) When $p > 1$ the energy decays as does $t^{-2/(p-1)}$.

It is important to note that the exponential decay rate of the energy may not be expected under the coercivity assumption (1.21) unless $p = 1$. The same phenomena appear in the nonlinear dissipative ordinary differential equation

$$u'' + u + |u'|^{p-1}u' = 0.$$

However, we are not now able to prove that the decay on (b) is optimal.

As mentioned above, the growth assumption (1.20) (which implies that the nonlinearity g is globally majorized by a linear function) is not necessary if we look for stability results asserting that every solution tends to the equilibrium state $\{0, 0\}$ in the energy space as $t \to \infty$ (see also [6] and [17]). However, this assumption will be needed to establish the a priori estimates leading to the decay rates. On the other hand, we note that the coerciveness assumption (1.21) concerns mainly the behavior of g at the origin and that the decay rate (of order $t^{-2/(p-1)}$) is governed by this behavior. This is a natural result since we already know that $\{y(t), y'(t)\} \to 0$, but it must be carefully proved since the decay holds only on the topology of $V \times L^2(\Omega)$.

It is easy to check that the hypotheses (1.20)-(1.21) are satisfied when g is, for instance, as follows:

$$g(s) = \begin{cases} |s|^{p-1}s & \text{when } |s| \leqq 1, \\ s & \text{when } |s| \geqq 1. \end{cases}$$

To handle the case where $p > 1$ in (1.21) we are forced to modify the perturbed energy functional introduced in [11] and [12]. This will be done by following the earlier works of Haraux and Zuazua [10] and Zuazua [23].

We note that in the situation considered above (where $\operatorname{int} \Gamma_*(x^0) \neq \varnothing$), the quantity $(E(t))^{1/2}$ is a norm in $V \times L^2(\Omega)$ equivalent to the one induced by $H^1(\Omega) \times L^2(\Omega)$. Therefore, it is equivalent to study the rate of decay of the energy or the rate of convergence of the solution to the equilibrium state $\{0, 0\}$ in the space $V \times L^2(\Omega)$. The situation is different when $\operatorname{int} \Gamma_*(x^0) = \varnothing$. In this case the quantity $(E(t))^{1/2}$ does not define a norm in $V \times L^2(\Omega) = H^1(\Omega) \times L^2(\Omega)$ and every constant function is a stationary solution of (1.3)-(1.6). In this case to recover the stability properties and the estimates on the rate of decay mentioned above we modify the feedback law and consider, instead of (1.4), (1.5), the following boundary condition:

$$\frac{\partial y}{\partial \nu} + \alpha\{m(x) \cdot \nu(x)\}y = -\{m(x) \cdot \nu(x)\}g(y') \quad \text{on } \Gamma(x^0) \times (0, \infty) \tag{1.22}$$

where $\alpha > 0$.

The rest of the paper is divided in two parts. In § 2 we give and prove our main result concerning the system (1.3)-(1.6) in the case where $\operatorname{int} \Gamma_*(x^0) \neq \varnothing$. In § 3 we give some remarks and discuss some possible extensions of this result and, in particular, the case where $\operatorname{int} \Gamma_*(x^0) = \varnothing$.

2. The main result. The main result of this paper is as follows.

THEOREM 2.1. *Let Ω be a bounded domain of $\mathbf{R}^n$, $n \geqq 3$, with smooth boundary $\Gamma = \partial\Omega$. Assume that $x^0 \in \mathbf{R}^n$ is such that* $\operatorname{int} \Gamma_*(x^0) \neq \varnothing$. *Let g be a continuous nondecreasing function such that $g(0) = 0$ and that (1.20)-(1.21) are satisfied for some positive constants C_1, $C_2 > 0$ and some $p \geqq 1$. Then we have*:

(a) *If $p = 1$, there exist some constants $M > 1$, $\delta > 0$ such that*

$$E(t) \leqq ME(0) \exp\{-\delta t\} \quad \forall t \geqq 0 \tag{2.1}$$

for any solution of (1.3)-(1.6).

(b) *If $p > 1$, there exist some constants $M > 1$ and $\mu > 0$ (with μ depending on $E(0)$) such that*

$$E(t) \leqq ME(0)(1+\mu t)^{-2/(p-1)} \quad \forall t \geqq 0 \tag{2.2}$$

for any solution of (1.3)-(1.6).

Proof. The proof will be carried out in several steps.

Step 1. The well-posedness of the problem is standard. The methods of Brezis [2], Haraux [9], and Lions and Magenes [19] may be applied (see also [17] for the study of this problem) to obtain the following lemma.

LEMMA 2.2. *Let us assume that the hypotheses of Theorem* 2.1 *are satisfied. Then, for any initial data $\{y^0, y^1\} \in V \times L^2(\Omega)$ there exists a uniquely weak solution $y = y(x, t)$ of* (1.3)-(1.6) *such that*

$$y \in C(\mathbf{R}^+, V) \cap C^1(\mathbf{R}^+; L^2(\Omega)), \quad E(t) \leqq E(0) \quad \forall t \geqq 0. \tag{2.3}$$

In addition, the following properties are verified:

(i) *Stability. If $\{y^0, y^1\}$ are replaced by $\{\hat{y}^0, \hat{y}^1\}$, then the corresponding solution $\hat{y}$ is such that*

$$\begin{aligned} &\frac{1}{2}\int_\Omega \{|\nabla y(x,t) - \nabla \hat{y}(x,t)|^2 + |y'(x,t) - \hat{y}'(x,t)|^2\}\, dx \\ &\qquad \leqq \frac{1}{2}\int_\Omega \{|\nabla y^0(x) - \nabla \hat{y}^0(x)|^2 + |y^1(x) - \hat{y}^1(x)|^2\}\, dx \quad \forall t \geqq 0. \end{aligned} \tag{2.4}$$

(ii) *Regularity. If in addition we assume $\{y^0, y^1\} \in W$ where*

$$W = \left\{\{y^0, y^1\} \in V \times V : \Delta y^0 \in L^2(\Omega); \frac{\partial y^0}{\partial \nu} = -\{m \cdot \nu\} g(y^1) \text{ on } \Gamma(x^0)\right\}, \tag{2.5}$$

then we have

$$y \in W^{1,\infty}(\mathbf{R}^+; V); \Delta y \in L^\infty(\mathbf{R}^+; L^2(\Omega)), \tag{2.6}$$

$$\frac{dE(t)}{dt} = E'(t) = -\int_{\Gamma(x^0)} \{m(x) \cdot \nu(x)\} g(y'(x,t)) y'(x,t)\, d\Gamma \quad \forall t \geqq 0. \tag{2.7}$$

From the stability property (2.4) and the fact that the set W is dense in $V \times L^2(\Omega)$ it is enough to prove the estimates (2.1), (2.2) for initial data $\{y^0, y^1\} \in W$, provided the constant μ depends continuously on $E(0)$. Therefore, in what follows we will consider the initial data in in W.

Step 2. We denote by $(\cdot, \cdot)$ (respectively, $|\cdot|$) the scalar product (respectively, the norm) in $L^2(\Omega)$. We introduce the functional

$$\rho(t) = 2(y'(t), m \cdot \nabla y(t)) + (n-1)(y'(t), y(t)). \tag{2.8}$$

We note that

$$|\rho(t)| \leqq C_3 E(t) \tag{2.9}$$

with $C_3 = 2R + (n-1)\alpha$, $\alpha > 0$ being the best constant such that

$$|\psi|_{L^2(\Omega)} \leqq \alpha |\nabla \psi|_{L^2(\Omega)} \quad \forall \psi \in V$$

and $R = \|m(x)\|_{L^\infty(\Omega)}$.

To simplify the notation we will omit the variable x of the functions under the integral sign.

Let us now calculate the derivative of the functional $\rho(t)$. We have (see [11], [12] for the details):

$$\begin{aligned} \rho'(t) &= 2(\Delta y(t), m \cdot \nabla y(t)) - (n-1)|\nabla y(t)|^2 - |y'(t)|^2 \\ &\quad + \int_{\Gamma(x^0)} \left[(n-1) \frac{\partial y(t)}{\partial \nu} y(t) + \{m \cdot v\} |y'(t)|^2 \right] d\Gamma. \end{aligned} \tag{2.10}$$

To estimate the first term of the right-hand side of (2.10) we need the following lemma, which is a slight generalization of an inequality due to Grisvard [8].

LEMMA 2.3. *Under the hypotheses of Theorem* 2.1 *the following inequality holds*:

$$2(\Delta y, m \cdot \nabla y) \leqq (n-2)|\nabla y|^2 + R^2 \int_{\Gamma(x^0)} \{m \cdot \nu\} |v|^2 \, d\Gamma \tag{2.11}$$

for every function $y \in V$ *such that* $\Delta y \in L^2(\Omega)$ *and* $\partial y / \partial \nu = \{m \cdot v\}$ v *with* $v \in L^2(\Gamma(x^0))$.

Proof of Lemma 2.3. In [8] inequality (2.11) has been proved in the case where $v \equiv 0$. Subsequently, in [11] and [12] the following inequality has been proved for $v \in H^{1/2}(\Gamma(x^0))$:

$$2(\Delta y, m \cdot \nabla y) \leqq (n-2)|\nabla y|^2 + 2 \int_\Gamma \frac{\partial y}{\partial \nu} m \cdot \nabla y \, d\Gamma - \int_\Gamma \{m \cdot \nu\} |\nabla y|^2 \, d\Gamma. \tag{2.12}$$

From (2.12) inequality (2.11) may be easily deduced [11], [12]. Let us recall the proof for the sake of completeness. We have

$$\begin{aligned} 2 \int_\Gamma \frac{\partial y}{\partial \nu} m \cdot \nabla y \, d\Gamma - \int_\Gamma \{m \cdot \nu\} |\nabla y|^2 \, d\Gamma &= 2 \int_{\Gamma(x^0)} \frac{\partial y}{\partial \nu} m \cdot \nabla y \, d\Gamma - \int_{\Gamma(x^0)} \{m \cdot \nu\} |\nabla y|^2 \, d\Gamma \\ &\quad + \int_{\Gamma_*(x^0)} \{m \cdot \nu\} |\nabla y|^2 \, d\Gamma \\ &\leqq 2 \int_{\Gamma(x^0)} \frac{\partial y}{\partial \nu} m \cdot \nabla y \, d\Gamma - \int_{\Gamma(x^0)} \{m \cdot \nu\} |\nabla y|^2 \, d\Gamma \\ &\leqq R^2 \int_{\Gamma(x^0)} \frac{1}{\{m \cdot \nu\}} \left| \frac{\partial y}{\partial \nu} \right|^2 d\Gamma \end{aligned}$$

since $y = 0$, $m \cdot \nu \leqq 0$ on $\Gamma_*(x^0)$.

Therefore we have proved (2.11) for $v \in H^{1/2}(\Gamma(x^0))$. This inequality may now be extended for all $v \in L^2(\Gamma(x^0))$ by a standard density argument. □

Combining (2.10) and Lemma 2.3, we deduce

$$\begin{aligned} \rho'(t) \leqq -2E(t) + \int_{\Gamma(x^0)} \{m \cdot \nu\} [R^2 |g(y'(t))|^2 - (n-1) g(y'(t)) y(t) \\ + |y'(t)|^2] \, d\Gamma \quad \forall t \geqq 0 \end{aligned} \tag{2.13}$$

and taking into account that

$$(n-1)\left|\int_{\Gamma(x^0)}\{m\cdot\nu\}g(y'(t))y(t)\,d\Gamma\right|\leqq E(t)+\frac{\beta(n-1)^2}{2}\int_{\Gamma(x^0)}\{m\cdot\nu\}|g(y'(t))|^2\,d\Gamma$$

where $\beta>0$ is the best constant such that

$$\int_{\Gamma(x^0)}\{m\cdot\nu\}|\psi|^2\,d\Gamma\leqq\beta|\nabla\psi|^2\quad\forall\psi\in V$$

and by applying (1.10), we deduce that

$$\rho'(t)\leqq -E(t)+C_4\int_{\Gamma(x^0)}\{m\cdot\nu\}|y'(t)|^2\,d\Gamma\quad\forall t\geqq 0 \tag{2.14}$$

for $C_4=(R^2+(\beta(n-1)^2)/2)C_1^2+1$.

Therefore from (2.3), (2.9), and (2.14) we deduce that

$$\begin{aligned}\frac{d[(E(t))^{(p-1)/2}\rho(t)]}{dt}&=\frac{p-1}{2}(E(t))^{(p-3)/2}E'(t)\rho(t)+(E(t))^{(p-1)/2}\rho'(t)\\ &\leqq -C_5E'(t)-(E(t))^{(p+1)/2}\\ &\quad+C_4(E(t))^{(p-1)/2}\int_{\Gamma(x^0)}\{m\cdot\nu\}|y'(t)|^2\,d\Gamma\quad\forall t\geqq 0\end{aligned} \tag{2.15}$$

with $C_5=((p-1)/2)C_3(E(0))^{(p-1)/2}$.

For $\varepsilon>0$ we introduce the functional

$$E_\varepsilon(t)=(1+\varepsilon C_5)E(t)+\varepsilon[E(t)]^{(p-1)/2}\rho(t). \tag{2.16}$$

The energy $E(t)$ being nonincreasing, we deduce that

$$\tfrac{1}{2}[E_\varepsilon(t)]^{(p+1)/2}\leqq[E(t)]^{(p+1)/2}\leqq 2[E_\varepsilon(t)]^{(p+1)/2}\quad\forall t\geqq 0 \tag{2.17}$$

provided we choose $\varepsilon>0$ such that

$$\varepsilon[E(0)]^{(p-1)/2}\leqq\frac{2}{C_3}\min\left\{\frac{1}{p+1}(2^{2/(p+1)}-1),\frac{1}{|p-3|}\left(1-\frac{1}{2^{2/(p+1)}}\right)\right\}. \tag{2.18}$$

From (2.7), (2.15) we deduce that

$$\begin{aligned}E'_\varepsilon(t)&\leqq -\int_{\Gamma(x^0)}\{m\cdot\nu\}g(y'(t))y'(t)\,d\Gamma\\ &\quad-\varepsilon(E(t))^{(p+1)/2}+\varepsilon C_4(E(t))^{(p-1)/2}\int_{\Gamma(x^0)}\{m\cdot\nu\}|y'(t)|^2\,d\Gamma\quad\forall t\geqq 0.\end{aligned} \tag{2.19}$$

By applying (1.21), we get

$$\begin{aligned}&\varepsilon C_4(E(t))^{(p-1)/2}\int_{\Gamma(x^0)\cap[|y'(t)|\geqq 1]}\{m\cdot\nu\}|y'(t)|^2\,d\Gamma\\ &\leqq\int_{\Gamma(x^0)\cap[|y'(t)|\geqq 1]}\{m\cdot\nu\}g(y'(t))y'(t)\,d\Gamma\quad\forall t\geqq 0\end{aligned} \tag{2.20}$$

for $\varepsilon>0$ such that

$$\varepsilon(E(0))^{(p-1)/2}C_4\leqq C_2. \tag{2.21}$$

Combining (2.19) and (2.20), we obtain

$$E'_\varepsilon(t) \leqq -\int_{\Gamma(x^0)\cap[|y'(t)|\leqq 1]} \{m\cdot\nu\} g(y'(t))y'(t)\,d\Gamma - \varepsilon(E(t))^{(p+1)/2} + \varepsilon C_4(E(t))^{(p-1)/2}\int_{\Gamma(x^0)\cap[|y'(t)|\leqq 1]} \{m\cdot\nu\}|y'(t)|^2\,d\Gamma \quad \forall t\geqq 0. \tag{2.22}$$

Step 3. We now distinguish the cases $p=1$ and $p>1$.

(a) Case $p=1$. In this case (1.21), (2.17), (2.21), and (2.22) imply

$$E'_\varepsilon(t) \leqq -\varepsilon E(t) \leqq -\frac{\varepsilon}{2}E_\varepsilon(t) \quad \forall t\geqq 0. \tag{2.23}$$

Solving inequality (2.23), we obtain

$$E_\varepsilon(t) \leqq E_\varepsilon(0)\exp\left\{-\frac{\varepsilon}{2}t\right\} \quad \forall t\geqq 0,$$

which combined with (2.17) yields

$$E(t) \leqq 4E(0)\exp\left\{-\frac{\varepsilon}{2}t\right\} \quad \forall t\geqq 0.$$

Taking into account that in this case the restrictions (2.18), (2.21) do not depend on $E(0)$, we obtain (2.1) with $M=4$ and $\delta=\varepsilon/2$. In fact, (2.1) may be proved for any $M>1$ by taking $\delta>0$ smaller.

(b) Case $p>1$. We apply Young's inequality

$$[E(t)]^{(p-1)/2}\int_{\Gamma(x^0)\cap[|y'(t)|\leqq 1]} \{m\cdot\nu\}|y'(t)|^2\,d\Gamma \leqq \lambda^{(p+1)/(p-1)}[E(t)]^{(p+1)/2} + \frac{1}{\lambda^{(p+1)/2}}\left(\int_{\Gamma(x^0)\cap[|y'(t)|\leqq 1]} \{m\cdot\nu\}|y'(t)|^2\,d\Gamma\right)^{(p+1)/2}$$

valid for all $\lambda>0$ with $\lambda=(1/2C_4)^{(p-1)/(p+1)}$, obtaining

$$E'_\varepsilon(t) \leqq -\int_{\Gamma(x^0)\cap[|y'(t)|\leqq 1]} \{m\cdot\nu\} g(y'(t))y'(t)\,d\Gamma - \frac{\varepsilon}{2}(E(t))^{(p+1)/2} + \varepsilon(2C_4)^{(p-1)/2}\left(\int_{\Gamma(x^0)\cap[|y'(t)|\leqq 1]} \{m\cdot\nu\}|y'(t)|^2\,d\Gamma\right)^{(p+1)/2} \quad \forall t\geqq 0. \tag{2.24}$$

By applying Hölder's inequality, we deduce

$$\left(\int_{\Gamma(x^0)\cap[|y'(t)|\leqq 1]} \{m\cdot\nu\}|y'(t)|^2\,d\Gamma\right)^{(p+1)/2} \leqq \left(\int_{\Gamma(x^0)\cap[|y'(t)|\leqq 1]} \{m\cdot\nu\}\,d\Gamma\right)^{(p-1)/2}\int_{\Gamma(x^0)\cap[|y'(t)|\leqq 1]} \{m\cdot\nu\}|y'(t)|^{p+1}\,d\Gamma \leqq \left(\int_{\Gamma(x^0)} \{m\cdot\nu\}\,d\Gamma\right)^{(p-1)/2}\int_{\Gamma(x^0)\cap[|y'(t)|\leqq 1]} \{m\cdot\nu\}|y'(t)|^{p+1}\,d\Gamma \quad \forall t\geqq 0,$$

which combined with (1.21), (2.24) for $\varepsilon>0$ small enough verifying

$$\varepsilon(2C_4)^{(p-1)/2}\left(\int_{\Gamma(x^0)} \{m\cdot\nu\}\,d\Gamma\right)^{(p-1)/2} \leqq C_2 \tag{2.25}$$

yields

$$E_\varepsilon'(t) \leqq -\frac{\varepsilon}{2}(E(t))^{(p+1)/2} \leqq -\frac{\varepsilon}{4}(E_\varepsilon(t))^{(p+1)/2} \quad \forall t \geqq 0. \tag{2.26}$$

Finally, solving inequality (2.26), we get

$$E_\varepsilon(t) \leqq E_\varepsilon(0)\left\{1+\frac{\varepsilon(p-1)}{8}[E_\varepsilon(0)]^{(p-1)/2}t\right\}^{-2/(p-1)} \quad \forall t \geqq 0$$

and then, from (2.17),

$$E(t) \leqq 4E(0)\left\{1+\frac{\varepsilon(p-1)}{2^{(4p+2)/p+1}}[E(0)]^{(p-1)/2}t\right\}^{-2/(p-1)} \quad \forall t \geqq 0. \tag{2.27}$$

The restrictions we have made on the choice of ε are (2.18), (2.21), (2.25), which show that, in (2.27), the quantity $\varepsilon[E(0)]^{(p-1)/2}$ may be chosen so that

$$\varepsilon[E(0)]^{(p-1)/2} = \min\,(C_6, C_7[E(0)]^{(p-1)/2}\}$$

for some positive constants C_6, C_7 that do not depend on $E(0)$. We see in particular, that the constant μ of (2.2) depends continuously on $E(0)$. The proof of Theorem 2.1 is now completed. □

Remark 2.4. The hypothesis $n \leqq 3$ has been implicitly used in the proof of Lemma 2.3 since we have applied Grisvard's inequality that has been proved only when $n \leqq 3$. We note that restriction $n \leqq 3$ is not needed when $\overline{\Gamma(x^0)} \cap \overline{\Gamma_*(x^0)} = \varnothing$. Indeed, in this case Lemma 2.3 may be easily proved for every n by applying Green's formula (see, for instance, [3], [13]).

Note also that the method proof of Theorem 2.1 is general and would apply to dimensions $n \geqq 4$ provided the natural generalization of Grisvard's inequality is proved. But it seems this has not been done.

Remark 2.5. The calculations leading to the estimates (2.1), (2.2) do not utilize the assumption that g is nondecreasing. Thus, these estimates also hold for every sufficiently smooth (to carry on the calculations above) solution of (1.3)-(1.6) even if g is not monotone. However, the stability property (2.4) is not satisfied unless g is nondecreasing and then estimates (2.1), (2.2) cannot be extended to weak solutions.

3. Further remarks.

3.1. The case where int $\Gamma_*(x^0) = \varnothing$. As has been pointed out in the Introduction, when int $\Gamma_*(x^0) = \varnothing$ (i.e., Ω is star-shaped with respect to x^0), the system (1.3)-(1.6) reduces to (1.3), (1.4), (1.6) and there exist nontrivial stationary solutions of it (every constant function is a solution). To obtain rates of decay on the $H^1(\Omega) \times L^2(\Omega)$-norm of solutions, we replace the boundary conditions (1.4), (1.5) by (1.22) with $\alpha > 0$. The energy associated to the system is then

$$E_\alpha(t) = \frac{1}{2}\int_\Omega \{|\nabla y(x,t)|^2 + |y'(x,t)|^2\}\,dx + \frac{\alpha}{2}\int_\Gamma \{m(x)\cdot\nu(x)\}|y(x,t)|^2\,d\Gamma. \tag{3.1}$$

We note that since $\alpha > 0$, $(E_\alpha(t))^{1/2}$ defines a norm in $V \times L^2(\Omega) = H^1(\Omega) \times L^2(\Omega)$ equivalent to the usual one. Then it is equivalent to obtain estimates for the rate of decay of $E_\alpha(t)$ or of the $H^1(\Omega) \times L^2(\Omega)$-norm.

We have the following result.

THEOREM 3.1. *Let Ω be a bounded domain of $\mathbf{R}^n$ with smooth boundary $\Gamma = \partial\Omega$ and star-shaped with respect to $x^0 \in \mathbf{R}^n$. Let g be a nondecreasing continuous function such that $g(0) = 0$ and that* (1.20)-(1.21) *are satisfied for some positive constants C_1,*

$C_2 > 0$ and some $p \geqq 1$. Then, if

$$\alpha \in \left(0, \frac{n}{2R^2}\right) \quad \text{with } R = \|x - x^0\|_{L^\infty(\Omega)} \tag{3.2}$$

the conclusions (2.1), (2.2) *of Theorem* 2.1 *hold for the energy $E_\alpha(t)$ defined in* (3.1) *for every solution of the system* (1.3), (1.6), (1.22).

Sketch of proof. First we note that the proof of Theorem 2.1 may also be done by choosing the following functional instead of $\rho(t)$:

$$\rho_\theta(t) = 2(y'(t), m \cdot \nabla y(t)) + \theta(y'(t), y(t)) \tag{3.3}$$

with $\theta \in (n-2, n)$.

Now let us fix $\theta \in (n-2, n)$ such that

$$\alpha < \frac{\theta}{2R^2} \tag{3.4}$$

(which is possible since (3.2) is satisfied) and calculate the derivative of the corresponding functional $\rho_\theta(t)$. By applying Lemma 2.3 we easily get

$$\begin{aligned}\rho_\theta'(t) &\leqq -(\theta - n + 2)|\nabla y(t)|^2 - (n-\theta)|y'(t)|^2 + \int_\Gamma \{m \cdot \nu\} \\ &\quad \times [2R^2|g(y'(t))|^2 - \theta g(y'(t))y(t) + |y'(t)|^2]\, d\Gamma - \alpha(\theta - 2R^2\alpha) \int_\Gamma \{m \cdot \nu\}|y(t)|^2\, d\Gamma \\ &\leqq -2\gamma E_\alpha(t) + \int_\Gamma \{m \cdot \nu\}[2R^2|g(y'(t))|^2 - \theta g(y'(t))y(t) + |y'(t)|^2]\, d\Gamma \quad \forall t \geqq 0\end{aligned}$$

with

$$\gamma = \min\{\theta - n + 2,\ n - \theta,\ \theta - 2R^2\alpha\}.$$

On the other hand, the energy $E_\alpha(t)$ is such that the identity (1.9) holds and the rest of the proof is analogous to that of Theorem 2.1. □

Remark 3.2. In the case where the function g is linear we know that the restriction (3.2) on α is unnecessary (see [21]). It would be interesting to study whether or not this hypothesis is necessary in the nonlinear framework.

Remark 3.3. The feedback law (1.22) with $\alpha > 0$ is more robust than (1.4) with respect to the perturbations of the support $\Gamma(x^0)$ of the boundary damping since the energy (3.1) is always coercive in the space $V \times L^2(\Omega)$ (see [22] for the study of these questions).

Remark 3.4. In the case where $\operatorname{int} \Gamma_*(x^0) \neq \varnothing$ and the boundary condition (1.22) is considered instead of (1.4) on $\Gamma(x^0)$, the conclusions of Theorem 2.1 remain valid provided $|\alpha|$ is sufficiently small. We note that, in this case, the restriction $\alpha > 0$ is not necessary but it is essential that $(E_\alpha(t))^{1/2}$ remains coercive in $V \times L^2(\Omega)$.

3.2. Nonhomogenous wave equation. Assume that $\operatorname{int} \Gamma_*(x^0) \neq \varnothing$ and let us consider the system

$$\begin{aligned} &y'' - \Delta y = h(x) && \text{in } \Omega \times (0, \infty), \\ &\frac{\partial y}{\partial v} = -\{m(x) \cdot \nu(x)\} g(y') && \text{on } \Gamma(x^0) \times (0, \infty), \\ &y = 0 && \text{on } \Gamma_*(x^0) \times (0, \infty), \\ &y(0) = y^0 \in V, \qquad y'(0) = y^1 \in L^2(\Omega) \end{aligned} \tag{3.5}$$

with $h(x) \in L^2(\Omega)$.

There exists a unique stationary solution $y^* \in V$ given by

$$
\begin{aligned}
&-\Delta y^* = h(x) && \text{in } \Omega, \\
&\frac{\partial y^*}{\partial \nu} = 0 && \text{on } \Gamma(x^0), \\
&y^* = 0 && \text{on } \Gamma_*(x^0).
\end{aligned} \tag{3.6}
$$

We observe that when $y = y(x, t)$ solves (3.5), then

$$z(x, t) = y(x, t) - y^*(x)$$

solves (1.3)-(1.5). Thus, Theorem 2.1 provides estimates on the rate of convergence in the space $V \times L^2(\Omega)$ of every solution of (3.5) to the unique rest point y^*. This remark remains valid in the situations considered in § 3.1 above.

It would be interesting to consider problems of type (3.5) with $h = h(x, t)$ periodic or almost-periodic functions (with respect to t) and to study, following [9], [10], and [23], the existence of periodic or almost-periodic solutions and to obtain estimates for the rate of decay of the energy of the difference of two solutions of the problem.

3.3. Semilinear wave equation. Let $f \in W^{1,\infty}_{\text{loc}}(\mathbf{R})$ be a locally Lipschitz continuous function and let us consider the semilinear wave equation

$$y'' - \Delta y + f(y) = 0 \quad \text{in } \Omega \times (0, \infty) \tag{3.7}$$

that we complete with the boundary and initial conditions (1.4)-(1.6) ((1.6), (1.22) when $\operatorname{int} \Gamma_*(x^0) = \varnothing$).

Let us assume that the nonlinear function f satisfies the following sign and growth assumptions:

$$f(s)s \geqq 0 \quad \forall s \in \mathbf{R}, \tag{3.8}$$

$$
\begin{aligned}
&\exists C > 0, \quad p > 1, \quad (n-2)p \leqq n: |f(s) - f(z)| \\
&\qquad \leqq C(1 + |s|^{p-1} + |z|^{p-1})|s - z| \quad \forall s, z \in \mathbf{R},
\end{aligned} \tag{3.9}
$$

$$\exists \delta > 0: f(s)s \geqq (2+\delta)F(s) \quad \forall s \in \mathbf{R} \quad \text{where } F(z) = \int_0^z f(s)\, ds. \tag{3.10}$$

The energy associated to the system is then (if $\alpha = 0$)

$$E(t) = \frac{1}{2} \int_\Omega \{|\nabla y(x, t)|^2 + |y'(x, t)|^2\}\, dx + \int_\Omega F(y(x, t))\, dx. \tag{3.11}$$

All the results of the preceding sections may be generalized (with minor modifications in the proofs) to obtain decay rates of type (2.1), (2.2) for the energy given by (3.11) (see [12] for a detailed proof in the case where g is linear).

3.4. More general partitions of the boundary. Let us return to system (1.3)-(1.6). In the case where g is linear it is well known that other partitions of the boundary (not necessarily of type $(\Gamma(x^0), \Gamma_*(x^0))$) give the exponential decay rate of the energy (see [1], [20] where very general results are proved by microlocal analysis techniques and [15] for a class of multipliers slightly larger than the radial one considered in this paper). It would be interesting to study whether, under those partitions, nonlinear feedback laws may be handled and decay rates of type (2.1), (2.2) obtained.

Acknowledgment. The author thanks Professor J. Lagnese for fruitful discussions and support.

REFERENCES

[1] C. BARDOS, G. LEBEAU, AND J. RAUCH, *Contrôle et stabilisation dans les problèmes hyperboliques*, Appendix 2 in Contrôlabilité exacte de systèmes distribués. Méthode HUM, Collection RMA, Vol. 8, Masson, Paris, 1988.

[2] H. BREZIS, *Maximal Monotone Operators*, North Holland, Amsterdam, 1973.

[3] G. CHEN, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pure. Appl., 58 (1979), pp. 249-274.

[4] ———, *Control and stabilization for the wave equation in a bounded domain* I, SIAM J. Control Optim., 17 (1979), pp. 66-81.

[4b] ———, *Control and stabilization for the wave equation in a bounded domain* II, SIAM J. Control Optim., 19 (1981), pp. 114-122.

[5] ———, *A note on the boundary stabilization of the wave equation*, SIAM J. Control Optim., 19 (1981), pp. 106-113.

[6] G. CHEN AND H. K. WANG, *Asymptotic behavior of solutions on the one-dimensional wave equation with a nonlinear elastic dissipative boundary condition*, to appear.

[7] R. DATKO, *Uniform asymptotic stability of evolutionary processes in Banach spaces*, SIAM J. Math. Anal., 3 (1972), pp. 428-445.

[8] P. GRISVARD, *Contrôlabilité exacte avec des conditions mêlées*, C. R. Acad. Sci. Paris Sér. I Math., 305 (1987), pp. 363-366.

[9] A. HARAUX, *Semilinear hyperbolic problems in bounded domains*, in Mathematical Reports, Vol. 3, J. Dieudonné, ed., Harwood Academic Publishers, Gordon and Breach, New York, 1987.

[10] A. HARAUX AND E. ZUAZUA, *Decay estimates for some semilinear damped hyperbolic problems*, Arch. Rational Mech. Anal., 100 (1988), pp. 191-206.

[11] V. KOMOMIK AND E. ZUAZUA, *Stabilisation frontière de l'equation des ondes: une méthode directe*, C. R. Acad. Sci. Paris Sér. I Math., 305 (1987), pp. 605-608.

[12] ———, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pure. Appl., to appear.

[13] J. LAGNESE, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163-182.

[14] ———, *Boundary stabilization of linear elastodynamic systems*, SIAM J. Control Optim., 21 (1983), pp. 968-984.

[15] ———, *Note on the boundary stabilization of wave equations*, SIAM J. Control Optim., 26 (1988), pp. 1250-1256.

[16] ———, *Boundary Stabilization of Thin Plates*, SIAM Studies in Applied Mathematics 10, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989.

[17] I. LASIECKA, *Stabilization of wave equations with nonlinear dissipative damping on the boundary*, in Proc. 26th IEEE Conference on Decision and Control, Los Angeles, CA, 1987, pp. 2348-2349.

[18] ———, *Stabilization of wave and plate-like equations with nonlinear dissipation on the boundary*, Applied Mathematics Report No. RM-88-05, University of Virginia, Charlottesville, VA, March 1988.

[19] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes*, Dunod, Paris, 1968.

[20] J. RAUCH AND M. E. TAYLOR, *Exponential decay of solutions to hyperbolic equations in bounded domains*, Indiana Univ. Math J., 24 (1974), pp. 79-86.

[21] E. ZUAZUA, *Some remarks on the boundary stabilizability of the wave equation*, in Control of Boundaries and Stabilization, J. Simon, ed., Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, New York, 1990.

[22] ———, *Robustesse du feedback de stabilisation par contrôle frontière*, C. R. Acad. Sci. Paris Ser. I Math., 307 (1988), pp. 587-591.

[23] ———, *Stability and decay for a class of nonlinear hyperbolic problems*, Asymptotic Anal., 1 (1988), pp. 161-185.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 478-481, March 1990

013

A NEW PROOF OF THE LYAPUNOV CONVEXITY THEOREM*

FABIO TARDELLA†

Abstract. A new proof of the Lyapunov Theorem is given, based on the Shapley-Folkman Theorem, that does not require any tools of functional analysis.

Key words. vector-valued measures, Lyapunov Convexity Theorem, convex analysis, integrals of multifunctions

AMS(MOS) subject classifications. 28B05, 46G10

1. Introduction. The Lyapunov Theorem appeared for the first time in [14]. It states that a nonatomic measure that takes values in $\mathbf{R}^n$ has closed and convex range. Most of its applications are to be found in the theory of optimal control and calculus of variations (see, e.g., [2], [8], [9], [18]). However it has also been fruitfully used in other fields, such as economics [16] and differential equations [19].

Because of its importance, many proofs of the Lyapunov Theorem have been given (see [3]-[5], [10]-[13], [20]). Nevertheless, the proofs appearing in the literature are rather involved or employ sophisticated theorems from functional analysis, such as the Krein-Milman Theorem, or compactness theorems in infinite-dimensional spaces.

Our purpose is to give a short proof of the Lyapunov Theorem that does not digress from measure theory and convex analysis. In fact the burden of the proof is carried by the Shapley-Folkman Theorem, a result of convex analysis (discovered by two economists) whose applications have unfortunately been largely restricted to mathematical economics [1], [17]. See, however, Appendix 1 of [7] for an interesting application to duality in mathematical programming.

SHAPLEY-FOLKMAN THEOREM. *Consider a finite family* $[C_i]_{i\in I}$ *of subsets of* $\mathbf{R}^n$. *If*

$$x \in \operatorname{co} \sum_{i\in I} C_i,$$

then there exists a subset J *of* I, *of cardinality at most* n, *such that*

$$x \in \sum_{i\notin J} C_i + \operatorname{co} \sum_{i\in J} C_i.$$

Here co S *denotes the convex hull of a set* $S \subset \mathbf{R}^n$.

As Blackwell has noted [3], the convexity of the range of a nonatomic vector measure is only a special case of a more general fact, namely, the convexity of the integral of a vector-valued multifunction. Since no extra work is required, we prove the latter statement in Theorem 1. The remaining part of the Lyapunov Theorem, namely, closedness, is established in Theorem 2.

We finally remark that the nonatomicity assumption is not only sufficient but, in a sense, also necessary for a vector measure to have a convex range.

2. The Lyapunov Theorem. Let (T, Σ, μ) be a positive measure space. The measure μ is said to be nonatomic if for every $A \in \Sigma$ with $\mu(A) > 0$, there is $B \in \Sigma$, $B \subset A$, such

* Received by the editors September 1, 1988; accepted for publication April 25, 1989.

† Istituto di Elaborazione dell'Informazione, C.N.R., Via Santa Maria 46, 56100 Pisa, Italy. This paper was written while the author was on leave at the C.R.M., University of Montreal, Canada.

that $0<\mu(B)<\mu(A)$. Since Σ is a σ-algebra, the nonatomicity of μ is easily shown to be equivalent to the following "Darboux property" [6, p. 25]:

$$\forall \alpha \in [0,1], \forall A \in \Sigma, \exists B \in \Sigma, B \subset A, \quad \text{such that } \mu(B) = \alpha\mu(A).$$

A function that associates with every t in T a subset $F(t)$ of $\mathbf{R}^n$ is called a multifunction from T to $\mathbf{R}^n$. An integrable function from T to $\mathbf{R}^n$ such that, for almost every t, $f(t) \in F(t)$ is called an integrable selection of F. The integral of F over T, denoted $\int_T F$, is defined as the set of all points in $\mathbf{R}^n$ of the form $\int_T f(t)\, d\mu$, where f is an integrable selection of F.

THEOREM 1. *If μ is finite and nonatomic, we have*

$$\int_T F = \operatorname{co} \int_T F.$$

Proof. We will show that if x_1 and x_2 belong to $\int_T F$, then the whole segment joining x_1 and x_2 is contained in $\int_T F$. Let $x_1 = \int_T f_1(t)\, d\mu$ and $x_2 = \int_T f_2(t)\, d\mu$, where f_1 and f_2 are integrable selections of F. We consider the multifunction $G(t) = \{f_1(t), f_2(t)\} \subset F(t)$. Because of the nonatomicity of μ, we can find a family $\{A_i\}_{1 \leqq i \leqq 2n}$ of elements of Σ such that $A_i \cap A_j = \varnothing$ for $i \neq j$, $\bigcup_{i=1,2n} A_i = T$, and $\mu(A_i) = (2n)^{-1}\mu(T)$ for $1 \leqq i \leqq 2n$. Given a point

$$x \in \operatorname{co} \int_T G = \operatorname{co} \sum_{i=1}^{2n} \int_{A_i} G$$

the Shapley–Folkman Theorem allows us to find a set I of n indices between 1 and $2n$, such that

$$x \in \sum_{i \notin I} \int_{A_i} G + \operatorname{co} \sum_{i \in I} \int_{A_i} G.$$

The above relation is equivalent to

$$x \in \int_{S_1} G + \operatorname{co} \int_{T \setminus S_1} G$$

where we have set $S_1 = \bigcup_{i \notin I} A_i$. We can therefore assert the existence of two points $x_1 \in \int_{S_1} G$ and $z_1 \in \operatorname{co} \int_{T \setminus S_1} G$ such that

$$x = x_1 + z_1 \quad \text{and} \quad \mu(S_1) = \mu(T \setminus S_1) = 2^{-1}\mu(T).$$

The same argument can be iteratively repeated replacing T by $T \setminus S_1$, so that, by induction, at the mth step we obtain m measurable sets $\{S_i\}_{1 \leqq i \leqq m}$ and $m+1$ points $\{x_i\}_{1 \leqq i \leqq m}$ and z_m, with $x_i \in \int_{S_i} G$ and $z_m \in \operatorname{co} \int_{T \setminus \bigcup_{i=1}^m S_i} G$ such that

$$x = \sum_{i=1}^{m} x_i + z_m \quad \text{and} \quad \mu(S_i) = 2^{-i}\mu(T), \quad 1 \leqq i \leqq m.$$

Observe now that $z_m \to 0$, since $\mu(T \setminus \bigcup_{i=1}^m S_i) = 2^{-m}\mu(T) \to 0$. Therefore we have

$$x = \sum_{i=1}^{\infty} x_i \in \int_{\bigcup_{i=1}^{\infty} S_i} G = \int_T G,$$

with the equality between the integrals holding because $\mu(\bigcup_{i=1}^{\infty} S_i) = \mu(T)$. We have thus shown that $\int_T G$ is convex. To complete the proof it is now sufficient to note that $x_1, x_2 \in \int_T G \subset \int_T F$. □

Let $m=(\mu_1, \mu_2, \cdots, \mu_n)$, where the μ_i are finite signed measures on the measurable space (T, Σ). We say that m is nonatomic if the total variation $|\mu_i|$ of every μ_i is nonatomic.

Remark 1. An equivalent way of defining the nonatomicity of a measure is that of requiring that for every $A\in\Sigma$ there is a $B\subset A$, $B\in\Sigma$, such that $\mu(B)\neq 0$ and $\mu(B)\neq\mu(A)$. This definition can also be applied to the more general case of a measure taking its values in a topological vector space.

We will denote by

$$R(\Sigma)=\{m(A)\colon A\in\Sigma\}$$

the range of the vector measure m, with respect to Σ, and by

$$\Sigma_A=\{A\cap B\colon B\in\Sigma\}$$

the trace of the σ-algebra Σ on a set $A\in\Sigma$.

The dimension of a subset C of $\mathbf{R}^n$, denoted dim C, is defined as the dimension of the smallest affine suspace of $\mathbf{R}^n$ containing C.

THEOREM 2 (Lyapunov Theorem). *If m is nonatomic, $R(\Sigma)$ is closed and convex.*

Proof. The measures $\mu_1, \mu_2, \cdots, \mu_n$ are all absolutely continuous with respect to the nonatomic measure $\mu=|\mu_1|+|\mu_2|+\cdots+|\mu_n|$. Then, by the Radon-Nikodym Theorem, there is an integrable function f from T to $\mathbf{R}^n$ such that $m(A)=\int_A f(t)\,d\mu$ for every $A\in\Sigma$. Let us consider the multifunction $F(t)=\{0, f(t)\}$. It is easy to see that the integrable selections of F are of the form $f\cdot\chi_A$, where χ_A is the characteristic function of $A\in\Sigma$. We then have

$$R(\Sigma)=\int_T F$$

and hence $R(\Sigma)$ is convex by Theorem 1.

Observe now that $R(\Sigma)$ is trivially closed when dim $R(\Sigma)=0$. We assume that it is closed when dim $R(\Sigma)\leqq n-1$ and we will prove that the same thing is true when dim $R(\Sigma)=n$. Assume there is a point $y\in\operatorname{cl} R(\Sigma)\setminus R(\Sigma)$. Then by a standard separation argument, we can find $p\in\mathbf{R}^n$ such that

$$p\cdot y=\sup\{p\cdot x\colon x\in R(\Sigma)\}.$$

Let us consider the sets

$$R_k=\{t\in T\colon p\cdot f(t)<-1/k\},\quad S_k=\{t\in T\colon p\cdot f(t)\geqq -1/k\},\quad k=1,2,\cdots.$$

Obviously,

$$R_k\cap S_k=\varnothing\quad\text{and}\quad R_k\cup S_k=T,\quad k=1,2,\cdots.$$

Let $\{A_n\}$ be a sequence of sets in Σ such that $m(A_n)$ converges to y. We then have

$$\lim_{n\to\infty}\mu(A_n\cap R_k)=0,\qquad k=1,2,\cdots.$$

Hence, for every k, we have

$$y=\lim_{n\to\infty}m(A_n)=\lim_{n\to\infty}\left(\int_{A_n\cap S_k}f(t)\,dt+\int_{A_n\cap R_k}f(t)\,dt\right)=\lim_{n\to\infty}m(A_n\cap S_k).$$

We can then find an increasing function $\sigma\colon\mathbf{N}\to\mathbf{N}$ such that, setting $B_n=A_{\sigma(n)}\cap S_n$, we have

$$\lim_{n\to\infty}m(B_n)=y.$$

Observe now that the sets B_n can be partitioned as follows:

$$B_n = B_n^+ \cup B_n^0 \cup B_n^-$$

where $B_n^+ \subset P = \{t \in T: p \cdot f(t) > 0\}$, $B_n^0 \subset Z = \{t \in T: p \cdot f(t) = 0\}$ and $B_n^- \subset N = \{t \in T: p \cdot f(t) < 0\}$. Since $B_n^- \subset \{t \in T: -1/n \leqq p \cdot f(t) < 0\}$, we have $\lim_{n\to\infty} m(B_n^-) = 0$. Furthermore, $\lim_{n\to\infty} m(B_n^+) = m(P)$. Hence

$$y = m(P) + \lim_{n\to\infty} m(B_n^0)$$

so that $y \in m(P) + \text{cl } R(\Sigma_Z)$. The proof is easily completed observing that dim $R(\Sigma_Z) <$ dim $R(\Sigma)$, and thus $R(\Sigma_Z)$ is closed by the induction hypothesis. □

Remark 2. It has been shown by Halmos [12] that $R(\Sigma)$ is actually closed even without the assumption of nonatomicity of m.

It can be easily observed that the nonatomicity assumption on m implies not only the convexity of $R(\Sigma)$ but also the convexity of $R(\Sigma_A)$ for every $A \in \Sigma$. Conversely, taking into account Remark 1, it is clear that if $R(\Sigma_A)$ is convex for every $A \in \Sigma$, the measure m is nonatomic. These remarks are summarized in the following proposition.

PROPOSITION 3. *The measure m is nonatomic if and only if* $R(\Sigma_A)$ *is convex for every* $A \in \Sigma$.

REFERENCES

[1] R. M. ANDERSON, *An elementary core equivalence theorem*, Econometrica, 46 (1978), pp. 1483-1487.

[2] G. ARONSSON, *Finite bang-bang controllability for certain non-linear systems*, Proc. Royal Soc. Edinburgh Sect. A, 77 (1977), pp. 137-149.

[3] D. BLACKWELL, *The range of certain vector integrals*, Proc. Amer. Math. Soc., 2 (1951), pp. 390-395.

[4] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Mathematics 580, Springer-Verlag, New York, 1977.

[5] L. CESARI, *Convexity of the range of certain integrals*, SIAM J. Control, 13 (1975), pp. 666-676.

[6] N. DINCULEANU, *Vector Measures*, Pergamon Press, Berlin, 1967.

[7] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[8] H. HALKIN, *Lyapunov's Theorem on the range of a vector measure and Pontryagin's Maximum Principle*, Arch. Rational Mech. Anal., 10 (1962), pp. 296-304.

[9] ———, *On the necessary condition for the optimal control of nonlinear systems*, J. Analyse Math., 12 (1964), pp. 1-82.

[10] ———, *On a generalization of a theorem of Lyapunov*, J. Math. Anal. Appl., 10 (1965), pp. 325-329.

[11] ———, *Some further generalizations of a theorem of Lyapunov*, Arch. Rational Mech. Anal., 17 (1964), pp. 272-277.

[12] P. R. HALMOS, *The range of a vector measure*, Bull. Amer. Math. Soc., 54 (1948), pp. 416-421.

[13] J. LINDENSTRAUSS, *A short proof of Lyapunov's convexity theorem*, J. Math. Mech., 15 (1966), pp. 971-972.

[14] A. LYAPUNOV, *Sur les fonctions-vecteurs complètement additives*, Bull. Acad. Sci. URSS Ser. Math., 4 (1940), pp. 465-478.

[15] H. RICHTER, *Verallgemeinerung eines in der Statistik benoetigten Satzes der Masstheorie*, Math. Annal., 150 (1963), pp. 85-90.

[16] B. SHITOVITZ, *Oligopoly in markets with a continuum of traders*, Econometrica, 41 (1973), pp. 467-501.

[17] R. M. STARR, *Quasi-equilibria in markets with non-convex preferences*, Econometrica, 17 (1969), pp. 25-38.

[18] H. J. SUSSMAN, *The bang-bang problem for certain control systems in* GL(n, R), SIAM J. Control, 10 (1970), pp. 470-476.

[19] G. TADMOR, *Functional differential equations of retarded and neutral type: analytic solutions and piecewise continuous control*, J. Differential Equations, 51 (1984), pp. 151-181.

[20] J. A. YORKE, *Another proof of the Lyapunov convexity theorem*, SIAM J. Control, 9 (1971), pp. 351-353.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 482-497, March 1990

014

NECESSARY AND SUFFICIENT CONDITIONS FOR LOCAL OPTIMALITY OF A PERIODIC PROCESS*

QINGHONG WANG† AND JASON L. SPEYER†

Abstract. The accessory minimization problem constructed from the second variation about a periodic extremal path with free period is investigated. Both necessary and sufficient conditions for a periodic path to be a weak local minimum are compactly established through a matrix inequality and the existence of a solution to the Riccati differential equation over the period. Extensions to the optimality conditions for the infinitely-repeated periodic processes are included. The existence of a real symmetric periodic solution to the Riccati differential equation is shown to be necessary for optimality, and this condition, plus some requirements on the eigenvalues of the monodromy matrix, imply sufficiency for optimality.

Key words. optimal control, periodic process, necessary and sufficient condition

AMS(MOS) subject classification. 49B10

1. Introduction. An autonomous optimal periodic control problem consists of minimizing an average cost subject to time-invariant dynamic equations and periodic boundary conditions [13]. In many interesting engineering systems, the solution to this problem is not a steady-state path, but rather a periodic path which gives best performance [9], [11]. In this paper, both necessary and sufficient conditions for weak, local optimality of a periodic path are developed, whereas only sufficiency was obtained in [3], [6], [13]. After formulating the problem in § 2, some second variation properties of the periodic control process are discussed. First, the accessory minimum problem is described in § 3. Then, a necessary condition is derived through the conjugate point condition related to a periodic process. It is shown in § 4 that this necessary condition is equivalent to the existence of a solution to the Riccati differential equation over a period. In § 5, by solving the accessory minimum problem, a necessary condition concerning the optimality, with respect to the period and the initial state, is established where a certain matrix is required to be positive semi-definite. In § 6, the results of previous sections are combined into a set of necessary and sufficient conditions. For sufficiency, the necessity of the nonnegativity of the matrix condition is strengthened to have certain positive definiteness properties. In this sense the difference between the necessary and sufficient conditions is minimal. However, the optimality of a single-period path does not imply that the repeated periodic path, which is obtained by repeating the orbit an arbitrary number of times, is also a minimum. If weak variations around these repeated orbits are considered, the necessary and sufficient conditions need to be strengthened. It is shown in § 7 that, as the number of repeated orbits becomes infinite, the existence of a periodic solution to the Riccati differential equation is necessary for optimality. Furthermore, sufficiency can be proved by requiring that the monodromy matrix has no eigenvalues on the unit circle except for a pair of unit eigenvalues coupled in the same Jordan box. A set of necessary and sufficient conditions for local optimality of an infinitely-repeated periodic process is summarized in § 8. These conditions somewhat weaken the sufficient conditions given in [13] and establish a close relationship between necessary and sufficient conditions. Conclusions are given in § 9.

* Received by the editors October 17, 1988; accepted for publication (in revised form) January 6, 1989.
† Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, Austin, Texas 78712. This work was sponsored by National Science Foundation grant ECS-8413475.

2. Formulation of the optimal periodic control problem. In this section, the problem of periodic control is formulated and the first-order necessary conditions for optimality are stated.

The optimal periodic control problem consists of minimizing the performance criterion

$$J(u(\cdot), x(0), \tau) = \frac{1}{\tau}\int_0^\tau L(x(t), u(t))\, dt \tag{1}$$

with respect to the period $\tau \in T = (0, \infty)$, the p-vector control function $u(\cdot) \in U = \{u(\cdot): T \to R^p$, piecewise continuous$\}$, and the initial condition of the state variable $x(0) \in R^n$, subject to the time-invariant dynamic constraint

$$\dot{x}(t) = f(x(t), u(t)) \tag{2}$$

and the periodic boundary condition

$$x(\tau) = x(0). \tag{3}$$

It is assumed that τ is the time when the periodic boundary condition is first met. The equilibrium solution, for which the second variational theory of [1], [2] applies, is excluded, and the problem where the periodic boundary conditions are met repeatedly is analyzed in § 7.

The following assumption is made on the problem described in (1) to (3).

Assumption 1. $f(\cdot, \cdot)$ and $L(\cdot, \cdot)$ and their derivatives through second order are continuous with respect to both arguments.

Let $\lambda_0 \in R$ and $\lambda(\cdot): T \to R^n$ be the Lagrange multipliers corresponding to the cost criterion (1) and the dynamics (2), and let H^* be the variational Hamiltonian defined as

$$H^*(x(t), u(t), \lambda(t), \lambda_0) = \lambda_0 L(x(t), u(t)) + \lambda^T(t) f(x(t), u(t)). \tag{4}$$

The minimum principle for an optimal periodic control process is given by the following first-order necessary condition [8].

PROPOSITION 1. *A necessary condition for* $(\tilde{u}(\cdot), \tilde{x}(0), \tau) \in U_{\text{int}} \times R^n \times T$ *being optimal for the problem described in* (1)-(3) *is that there exist Lagrange multipliers* $\lambda_0 \geqq 0$ *and* λ, *not vanishing simultaneously, such that*

$$\dot{\tilde{x}} = f(\tilde{x}, \tilde{u}), \quad \dot{\lambda} = -H_x^{*T}(\tilde{x}, \tilde{u}, \lambda, \lambda_0), \quad 0 = H_u^*(\tilde{x}, \tilde{u}, \lambda, \lambda_0) \tag{5}$$

and

$$\tilde{x}(\tau) = \tilde{x}(0), \quad \lambda(\tau) = \lambda(0), \quad H^*(\tilde{x}(\tau), \tilde{u}(\tau), \lambda(\tau), \lambda_0) = \lambda_0 J(\tilde{u}(\cdot), \tilde{x}(0), \tau) \tag{6}$$

where U_{int} *denotes the interior of the set* U.

Note that subscripts with respect to x and u denote partial differentiations. The explicit arguments of the functions will be dropped when the presentation appears clear. The existence of $(\tilde{u}(\cdot), \tilde{x}(0), \tau) \in U_{\text{int}} \times R^n \times T$ satisfying (5) and (6) is assumed, and the pair $(\tilde{x}(t), \tilde{u}(t))$ is called an extremal of the problem.

Assumption 2. $(\tilde{x}(t), \tilde{u}(t))$ is normal [15], and λ_0 is normalized to unity, i.e., $\lambda_0 = 1$.

The variational Hamiltonian associated with $\lambda_0 = 1$ is defined as

$$H(x(t), u(t), \lambda(t)) \triangleq H^*(x(t), u(t), \lambda(t), 1) \tag{7}$$

and the augmented performance criterion $\bar{J}$ is

$$\bar{J}(u(\cdot), x(0), \tau) = \frac{1}{\tau}\int_0^\tau [H(x, u, \lambda) - \lambda^T \dot{x}]\, dt. \tag{8}$$

The following assumptions on the extremal solutions obtained from (5) and (6) are required for our consideration of the weak second variation.

Assumption 3. The strong form of the Legendre-Clebsch condition, $H_{uu} > 0$, is satisfied along the extremal path.

Assumption 4. (f_x, f_u) is completely controllable along the extremal path.

Remark. Assumption 4 implies Assumption 2 and strong normality as defined in [15].

The following definitions are given in order to simplify the notation:

$$\tilde{f}(t) \triangleq f(\tilde{x}(t), \tilde{u}(t)) \tag{9}$$

$$\tilde{H}(t) \triangleq H(\tilde{x}(t), \tilde{u}(t), \lambda(t)). \tag{10}$$

Similarly, $\tilde{H}_x(t)$, $\tilde{H}_u(t)$, $\tilde{f}_x(t)$, $\tilde{H}_{xx}(t)$, $\tilde{H}_{uu}(t)$, $\tilde{H}_{xu}(t)$, and $\tilde{H}_{ux}(t)$ are partial derivatives of $f(x, u)$ and $H(x, u, \lambda)$ evaluated on the nominal path. For example $\tilde{H}_x(t)$ is defined as

$$\tilde{H}_x(t) \triangleq H_x(\tilde{x}(t), \tilde{u}(t), \lambda(t)). \tag{11}$$

3. Second variation and the accessory minimum problem. The second variational cost associated with perturbations away from an extremal path is derived in [13] as

$$\begin{aligned} d^2\bar{J} = \frac{1}{\tau}\Bigg\{ & [\delta x^T(0) \quad d\tau] \begin{bmatrix} 0 & \tilde{H}_x^T \\ \tilde{H}_x & -\tilde{H}_x\tilde{f} \end{bmatrix}_{t=0} \begin{bmatrix} \delta x(0) \\ d\tau \end{bmatrix} \\ & + \int_0^\tau [\delta x^T \quad \delta u^T] \begin{bmatrix} \tilde{H}_{xx} & \tilde{H}_{xu} \\ \tilde{H}_{ux} & \tilde{H}_{uu} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta u \end{bmatrix} dt \Bigg\} \end{aligned} \tag{12}$$

where $\delta x(t) \triangleq x(t) - \tilde{x}(t)$, $\delta u(t) \triangleq u(t) - \tilde{u}(t)$. The norms of δx and δu are defined as $\|\delta x\| = \max_{t\in[0,T]} (\sum_{i=1}^n (\delta x_i(t))^2)^{1/2}$ and $\|\delta u\| = \max_{t\in[0,T]} (\sum_{i=1}^p (\delta u_i(t))^2)^{1/2}$. Assume that $\|\delta x\|$, $\|\delta u\|$, and $|d\tau|$ are sufficiently small so that the dynamics of the system can be approximated by the linearized differential equation

$$\delta\dot{x} = \tilde{f}_x \delta x + \tilde{f}_u \delta u, \tag{13}$$

and the linearized periodic boundary condition

$$\delta x(\tau) = \delta x(0) - \tilde{f}(\tau)\, d\tau. \tag{14}$$

In the accessory minimum problem, the quadratic cost (12) is minimized subject to the variational constraints (13) and (14). Let $\delta\lambda$ be the Lagrange multiplier associated with the dynamic equation (13), then $\bar{H}$ is the variational Hamiltonian associated with the accessory minimum problem defined as

$$\bar{H} = \frac{1}{2}[\delta x^T \quad \delta u^T] \begin{bmatrix} \tilde{H}_{xx} & \tilde{H}_{xu} \\ \tilde{H}_{ux} & \tilde{H}_{uu} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta u \end{bmatrix} + \delta\lambda^T[\tilde{f}_x \delta x + \tilde{f}_u \delta u]. \tag{15}$$

The first-order necessary conditions, or Euler-Lagrange equations, are

$$0 = \tilde{H}_{ux}\delta x + \tilde{H}_{uu}\delta u + \tilde{f}_u^T \delta\lambda \tag{16}$$

$$\delta\dot{\lambda} = -\tilde{H}_{xx}\delta x - \tilde{H}_{xu}\delta u - \tilde{f}_x^T \delta\lambda. \tag{17}$$

Since $\tilde{H}_{uu} > 0$ by Assumption 3, δu can be solved from (16) in terms of δx and $\delta\lambda$ as

$$\delta u = -\tilde{H}_{uu}^{-1}(\tilde{H}_{ux}\delta x + \tilde{f}_u^T \delta\lambda). \tag{18}$$

By substituting δu of (18) into (13) and (17), the Hamiltonian system of differential equations is obtained as

$$\begin{bmatrix} \delta\dot{x} \\ \delta\dot{\lambda} \end{bmatrix} = \begin{bmatrix} A(t) & -B(t) \\ -C(t) & -A^T(t) \end{bmatrix} \begin{bmatrix} \delta x \\ \delta\lambda \end{bmatrix} \tag{19}$$

where

$$A(t) = \tilde{f}_x(t) - \tilde{f}_u(t)\tilde{H}_{uu}^{-1}(t)\tilde{H}_{ux}(t) \tag{20}$$

$$B(t) = \tilde{f}_u(t)\tilde{H}_{uu}^{-1}(t)\tilde{f}_u^T(t) \tag{21}$$

$$C(t) = \tilde{H}_{xx}(t) - \tilde{H}_{xu}(t)\tilde{H}_{uu}^{-1}(t)\tilde{H}_{ux}(t). \tag{22}$$

The state transition matrix associated with (19) is

$$\phi(t, 0) \triangleq \begin{bmatrix} \phi_{11}(t, 0) & \phi_{12}(t, 0) \\ \phi_{21}(t, 0) & \phi_{22}(t, 0) \end{bmatrix},$$

where $\phi_{ij}(t, 0)$, $i, j = 1, 2$, are $n \times n$ matrices which are partitioned blocks of $\phi(t, 0)$. $\phi(t, 0)$ is propagated by

$$\dot{\phi}(t, 0) = \begin{bmatrix} A(t) & -B(t) \\ -C(t) & -A^T(t) \end{bmatrix} \phi(t, 0), \qquad \phi(0, 0) = I. \tag{23}$$

In the next two sections, the positivity of the second variation with respect to δu, $\delta x(0)$, and $d\tau$ is discussed. The optimality of δu for the accessory minimum problem obtained from (18) is discussed through a conjugate point condition, and then the accessory minimum problem with respect to $\delta x(0)$ and $d\tau$ is considered. Note that the starting point on the path for a periodic process is irrelevant. If the starting point is indexed by t_0, then $t_0 = 0$ in the above formulation can be replaced by any $t_0 \in [0, \tau)$, and the time interval $[0, \tau]$ can be replaced by $[t_0, t_0 + \tau]$.

4. Conjugate point condition of a periodic path. The conjugate point of an extremal path is related to the existence of a nonzero solution to (19) with the zero variations in the initial state and final constraint [4], [10], [15]. If this occurs, a conjugate path can be found which has the same cost as the extremal path. Since the starting point on a periodic path is not unique, the initial and final state variations can be simply written as $\delta x(t_0) = 0$ and $\delta x(t_0 + \tau) = 0$. The conjugate point of a periodic path can be characterized by the property described below.

DEFINITION. t' and t'' are mutually conjugate if there exists a nontrivial solution to (19) on $[t', t'']$ such that $\delta x(t') = \delta x(t'') = 0$, and $\delta x(t) \neq 0$ on $t \in (t', t'')$, where $t' < t'' \leqq t' + \tau$ [10].

By using the state transition matrix defined in (23), $\delta x(t')$ and $\delta x(t'')$ are related by

$$\delta x(t'') = \phi_{11}(t'', t')\delta x(t') + \phi_{12}(t'', t')\delta\lambda(t'). \tag{24}$$

A nontrivial solution exists for $\delta x(t') = \delta x(t'') = 0$ if and only if $\phi_{12}(t'', t')$ is not invertible. Therefore, if $\phi_{12}(t'', t')$ is not invertible, t' and t'' are mutually conjugate. For convenience, let $t' = t_0$ be named as the starting time and $t'' = t_c$ be named as the conjugate time. It will be shown that if there exists a conjugate time $t_c \in (t_0, t_0 + \tau]$, the extremal periodic path is not a minimum. The method to be used is to compare the second variational cost of the conjugate path and that of some nonextremal path. In order to do that, some assumptions are made.

Assumption 5. $\mathcal{N}[\phi_{12}(t_c, t_0)] \cap \mathcal{N}[B^{1/2}(t_c)\phi_{22}(t_c, t_0)] = \{0\}$, where $\mathcal{N}[\cdot]$ represents the null-space of the matrix.

Assumption 6. $\mathcal{N}[\phi_{12}(t_0+\tau, t_0)] \cap \mathcal{N}[B^{1/2}(t_0)(I-\phi_{22}(t_0+\tau, t_0))] = \{0\}$.

Assumption 5 implies that the velocity of the extremal path and the velocity of the conjugate path are different at the conjugate time t_c. Assumption 6 implies that if the conjugate time is at the end of the period, i.e., $t_c = t_0+\tau$, the velocity of the conjugate path at the initial time t_0 is different from the velocity of the conjugate path at the conjugate time $t_0+\tau$. In the following discussion, the initial state variation $\delta x(t_0)$ is assumed to be zero, and the variation of the period $d\tau$ is assumed to be zero as well. The optimality with respect to the variation in the initial state and the period will be discussed in the next section.

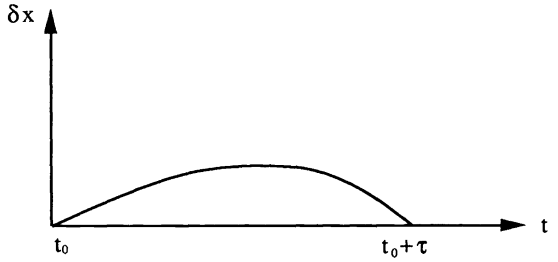
PROPOSITION 2. *If $\phi_{12}(t_c, t_0)$ is not invertible for some $t_c \in (t_0, t_0+\tau]$, then the extremal periodic path is not a minimum.*

Proof. The nonoptimality of the second variation due to a conjugate time $t_c < t_0+\tau$ has been proved elsewhere [4], [15, Thm. 3.1]. Although the case $t_c = t_0+\tau$ is usually excluded, for the periodic control problem, t_0 and t_c can correspond to the same point on the periodic path. By taking advantage of the closure of the path, some δu can be found to make the second variation negative. Let $\delta x(t_0) = 0$ and $\delta\lambda(t_0) = \beta$, where β is a nonzero vector in the null-space of $\phi_{12}(t_0+\tau, t_0)$. Define Path 1 and Path 2 in Fig. 1 to be the paths generated by $\delta u(t) = \delta u^1(t)$ and $\delta u(t) = \delta u^2(t)$, respectively, where

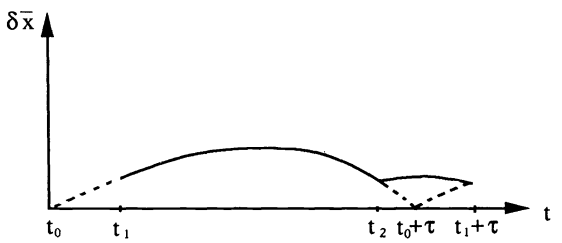
$$\delta u^1(t) = -\tilde{H}_{uu}^{-1}(\tilde{H}_{ux}\delta x + \tilde{f}_u^T \delta\lambda) \qquad t_0 \leqq t < t_0+\tau \tag{25}$$

$$\delta u^2(t) = \begin{cases} -\tilde{H}_{uu}^{-1}(\tilde{H}_{ux}\delta\bar{x} + \tilde{f}_u^T \delta\bar{\lambda}) & t_1 \leqq t \leqq t_2 \\ \text{extremal from} & \\ \delta\bar{x}(t_2) = \delta x(t_2) \text{ to } \delta\bar{x}(t_1+\tau) = \delta x(t_1) & t_2 < t < t_1+\tau \end{cases} \tag{26}$$

where $t_1 = t_0+\varepsilon_1$, $t_2 = t_0+\tau-\varepsilon_2$, and ε_1 and ε_2 are small positive numbers. Path 1 corresponds to the variation of the conjugate path, and Path 2 corresponds to the variation of the nonextremal path.



Path 1: Conjugate Path



Path 2: Nonextremal Path

FIG. 1. *Conjugate path and nonextremal path for $t_c = t_0+\tau$.*

If $\delta x(t)$ and $\delta\lambda(t)$ is a continuous solution of (19) on the interval (t_i, t_j), then by adding the zero term

$$
\begin{aligned}
0 &= -\int_{t_i}^{t_j} \delta\lambda^T[\delta\dot{x} - A(t)\delta x + B(t)\delta\lambda]\,dt \\
&= \delta\lambda^T(t_i)\delta x(t_i) - \delta\lambda^T(t_j)\delta x(t_j) \\
&\quad + \int_{t_i}^{t_j} [\delta\dot{\lambda}^T\delta x + \delta\lambda^T A(t)\delta x - \delta\lambda^T B(t)\delta\lambda]\,dt
\end{aligned} \tag{27}
$$

to the integral term

$$
\int_{t_i}^{t_j} [\delta x^T \quad \delta u^T]\begin{bmatrix} \tilde{H}_{xx} & \tilde{H}_{xu} \\ \tilde{H}_{ux} & \tilde{H}_{uu} \end{bmatrix}\begin{bmatrix} \delta x \\ \delta u \end{bmatrix} dt,
$$

and by substituting δu of (18) into the formulation, the following is obtained:

$$
\int_{t_i}^{t_j} [\delta x^T \quad \delta u^T]\begin{bmatrix} \tilde{H}_{xx} & \tilde{H}_{xu} \\ \tilde{H}_{ux} & \tilde{H}_{uu} \end{bmatrix}\begin{bmatrix} \delta x \\ \delta u \end{bmatrix} dt = \delta x^T(t_i)\delta\lambda(t_i) - \delta x^T(t_j)\delta\lambda(t_j). \tag{28}
$$

Let $J^k(t_i, t_j)$ be the second variational cost of the kth path on the interval (t_i, t_j). The cost of Path 1 is then

$$
J^1(t_0, t_0+\tau) = \frac{1}{\tau}\{\delta x^T(t_0)\delta\lambda(t_0) - \delta x^T(t_0+\tau)\delta\lambda(t_0+\tau)\} = 0. \tag{29}
$$

Our objective is to show that $J^2(t_1, t_1+\tau) < J^1(t_0, t_0+\tau)$. However, only $J^2(t_2, t_1+\tau) < J^1(t_0, t_1) + J^1(t_2, t_0+\tau)$ needs to be shown. $\delta x(t_i)$ and $\delta\lambda(t_i)$, $i = 1, 2$, can be written as

$$
\delta x(t_1) = \phi_{12}(t_1, t_0)\beta = -\varepsilon_1 B(t_0)\beta + o(\varepsilon_1) \tag{30}
$$

$$
\delta\lambda(t_1) = \phi_{22}(t_1, t_0)\beta = [I - \varepsilon_1 A^T(t_0)]\beta + o(\varepsilon_1) \tag{31}
$$

$$
\delta x(t_2) = \phi_{12}(t_2, t_0)\beta = \varepsilon_2 B(t_0)\phi_{22}(t_0+\tau, t_0)\beta + o(\varepsilon_2) \tag{32}
$$

$$
\delta\lambda(t_2) = \phi_{22}(t_2, t_0)\beta = [I + \varepsilon_2 A^T(t_0)]\phi_{22}(t_0+\tau, t_0)\beta + o(\varepsilon_2). \tag{33}
$$

By using (28) along with (30) to (33),

$$
\begin{aligned}
J^1(t_0, t_1) + J^1(t_2, t_0+\tau) &= \frac{1}{\tau}[-\delta x^T(t_1)\delta\lambda(t_1) + \delta x^T(t_2)\delta\lambda(t_2)] \\
&= \frac{1}{\tau}\beta^T[\varepsilon_1 B(t_0) + \varepsilon_2\phi_{22}^T(t_0+\tau, t_0)B(t_0)\phi_{22}(t_0+\tau, t_0)]\beta \\
&\quad + o(\varepsilon_1+\varepsilon_2).
\end{aligned} \tag{34}
$$

By using $\delta\bar{x}(t_2) = \delta x(t_2)$ and $\delta\bar{x}(t_1+\tau) = \delta x(t_1)$, the following can be obtained:

$$
\delta\bar{x}(t_2) = \delta x(t_2) = \varepsilon_2 B(t_0)\phi_{22}(t_0+\tau, t_0)\beta + o(\varepsilon_2) \tag{35}
$$

$$
\delta\bar{x}(t_1+\tau) = \delta x(t_1) = -\varepsilon_1 B(t_0)\beta + o(\varepsilon_1) \tag{36}
$$

$$
\begin{aligned}
\delta\bar{\lambda}(t_1+\tau) &= \phi_{21}(t_1+\tau, t_2)\delta\bar{x}(t_2) + \phi_{22}(t_1+\tau, t_2)\delta\bar{\lambda}(t_2) \\
&= -(\varepsilon_1+\varepsilon_2)C(t_0)\delta\bar{x}(t_2) + [I - (\varepsilon_1+\varepsilon_2)A^T(t_0)]\delta\bar{\lambda}(t_2) + o(\varepsilon_1+\varepsilon_2).
\end{aligned} \tag{37}
$$

$J^2(t_2, t_1+\tau)$ then can be written as

$$
\begin{aligned}
J^2(t_2, t_1+\tau) &= \frac{1}{\tau}[\delta\bar{x}^T(t_2)\delta\bar{\lambda}(t_2) - \delta\bar{x}^T(t_1+\tau)\delta\bar{\lambda}(t_1+\tau)] \\
&= \frac{1}{\tau}\beta^T[\varepsilon_2\phi_{22}^T(t_0+\tau, t_0)B(t_0)\delta\bar{\lambda}(t_2) + \varepsilon_1 B(t_0)\delta\bar{\lambda}(t_2)] + o(\varepsilon_1+\varepsilon_2).
\end{aligned} \tag{38}
$$

Since

$$\begin{aligned}\delta\bar{x}(t_1+\tau) &= \phi_{11}(t_1+\tau, t_2)\delta\bar{x}(t_2)+\phi_{12}(t_1+\tau, t_2)\delta\bar{\lambda}(t_2)\\ &= [I+(\varepsilon_1+\varepsilon_2)A(t_0)]\delta\bar{x}(t_2)-(\varepsilon_1+\varepsilon_2)B(t_0)\delta\bar{\lambda}(t_2)+o(\varepsilon_1+\varepsilon_2),\end{aligned} \tag{39}$$

by combining (35), (36), and (39),

$$(\varepsilon_1+\varepsilon_2)B(t_0)\delta\bar{\lambda}(t_2) = \varepsilon_2 B(t_0)\phi_{22}(t_0+\tau, t_0)\beta+\varepsilon_1 B(t_0)\beta+o(\varepsilon_1+\varepsilon_2). \tag{40}$$

Equation (38) becomes

$$\begin{aligned}J^2(t_2, t_1+\tau) &= \frac{1}{\tau}\frac{\varepsilon_2^2}{\varepsilon_1+\varepsilon_2}\beta^T\phi_{22}^T(t_0+\tau, t_0)B(t_0)\phi_{22}(t_0+\tau, t_0)\beta\\ &\quad+\frac{1}{\tau}\frac{1}{\varepsilon_1+\varepsilon_2}\beta^T[2\varepsilon_1\varepsilon_2 B(t_0)\phi_{22}(t_0+\tau, t_0)+\varepsilon_1^2 B(t_0)]\beta\\ &\quad+o(\varepsilon_1+\varepsilon_2).\end{aligned} \tag{41}$$

From Assumption 5, the first-order terms in (34) and (41) are nonzero. When ε_1 and ε_2 are small enough, the higher-order terms are ignored.

Define

$$\zeta = B^{1/2}(t_0)\beta, \qquad \eta = B^{1/2}(t_0)\phi_{22}(t_0+\tau, t_0)\beta. \tag{42}$$

Equations (34) and (41) can be written as

$$J^1(t_0, t_1)+J^1(t_2, t_0+\tau) = \frac{1}{\tau}(\varepsilon_2\eta^T\eta+\varepsilon_1\zeta^T\zeta) \tag{43}$$

$$J^2(t_2, t_1+\tau) = \frac{1}{\tau}\frac{1}{\varepsilon_1+\varepsilon_2}(\varepsilon_2^2\eta^T\eta+2\varepsilon_1\varepsilon_2\eta^T\zeta+\varepsilon_1^2\zeta^T\zeta). \tag{44}$$

From Assumption 6, $\zeta \neq \eta$. Therefore,

$$\eta^T\eta+\zeta^T\zeta > 2\eta^T\zeta \tag{45}$$

and

$$\begin{aligned}J^1(t_0, t_1)+J^1(t_2, t_0+\tau) &= \frac{1}{\tau}\frac{1}{\varepsilon_1+\varepsilon_2}(\varepsilon_1+\varepsilon_2)(\varepsilon_2\eta^T\eta+\varepsilon_1\zeta^T\zeta)\\ &= \frac{1}{\tau}\frac{1}{\varepsilon_1+\varepsilon_2}[\varepsilon_2^2\eta^T\eta+\varepsilon_1\varepsilon_2(\eta^T\eta+\zeta^T\zeta)+\varepsilon_1^2\zeta^T\zeta]\\ &> \frac{1}{\tau}\frac{1}{\varepsilon_1+\varepsilon_2}[\varepsilon_2^2\eta^T\eta+2\varepsilon_1\varepsilon_2\eta^T\zeta+\varepsilon_1^2\zeta^T\zeta]\\ &= J^2(t_2, t_1+\tau).\end{aligned} \tag{46}$$

This implies that the cost of the nonextremal path is less than the cost of the conjugate path. Therefore, the extremal periodic path is not a minimum. □

The existence of a conjugate point is closely related to the existence of a solution to the Riccati differential equation

$$\dot{P}(t) = -P(t)A(t)-A^T(t)P(t)+P(t)B(t)P(t)-C(t). \tag{47}$$

The following proposition establishes the relationship between these two.

PROPOSITION 3. *Given Assumptions* 1-4, *the necessary and sufficient condition for the absence of a point* $t_c \in (t_0, t_0+\tau]$ *conjugate to* t_0 *is that there exists a continuous real symmetric solution to the Riccati differential equation* (47) *on* $t \in [t_0, t_0+\tau]$.

Proof. For the proof, see the proof of Theorem 7.1 of [10].

COROLLARY 1. *A necessary condition for a periodic path to be a minimum is that there exists a continuous real symmetric solution* $P(t)$ *to the Riccati differential equation* (47) *on* $t \in [t_0, t_0+\tau]$.

COROLLARY 2. *If for every* $t_0 \in [0, \tau)$ *there exists a continuous real symmetric solution* $P(t)$ *to the Riccati differential equation* (47) *on the interval* $t \in [t_0, t_0+\tau]$, *then there are no mutual conjugate points along the extremal single-period path.*

Corollary 1 and Corollary 2 are immediate results of Propositions 2 and 3.

5. A local optimality condition for the initial state and period. If there is no conjugate point along the extremal path (see § 4), the solution to the accessory minimum problem can be determined for a given $\delta x(0)$ and $d\tau$ [5]. Define the quadratic cost $d^2\bar{J}^*(\delta x(0), d\tau)$ to be $\min_{\delta u(\cdot)} d^2\bar{J}$. Then, $d^2\bar{J}^*(\delta x(0), d\tau)$ is the solution to the accessory minimum problem calculated by finding a control $\delta u(\cdot)$ satisfying (13) and (14) which minimizes the second variational cost $d^2\bar{J}$. $d^2\bar{J}^*(\delta x(0), d\tau) \geqq 0$ is a necessary condition for optimality. Otherwise there exists a periodic path with period $\tau + d\tau$ and initial state $\tilde{x}(0) + \delta x(0)$ which gives less cost.

By defining the matrix $\bar{M}$ as

$$\bar{M}(0,\tau) = \begin{bmatrix} \phi_{12}^{-1}(I-\phi_{11}) + \phi_{12}^{-T}(I-\phi_{22}^{T}) & \tilde{H}_x^T + (\phi_{22}-I)\phi_{12}^{-1}\tilde{f} \\ \tilde{H}_x + \tilde{f}^T\phi_{12}^{-T}(\phi_{22}^T - I) & -\tilde{H}_x\tilde{f} - \tilde{f}^T\phi_{22}\phi_{12}^{-1}\tilde{f} \end{bmatrix}_{t=\tau} \tag{48}$$

where ϕ_{ij}, $i,j = 1, 2$, are partitioned blocks of $\phi(\tau, 0)$, and $\tilde{H}_x$ and $\tilde{f}$ are evaluated at $t = \tau$, a necessary condition for the optimality of a periodic process is stated below.

PROPOSITION 4. *Given that* $\phi_{12}(\tau, 0)$ *is invertible, a necessary condition for the periodic path to be a weak local minimum is* $\bar{M}(0, \tau)$ *being positive semi-definite.*

Proof. By using (28) in (12), $d^2\bar{J}^*(\delta x(0), d\tau)$ can be written as

$$\begin{aligned} d^2\bar{J}^*(\delta x(0), d\tau) = \frac{1}{\tau}\Bigg\{ & [\delta x^T(0) \quad d\tau] \begin{bmatrix} 0 & \tilde{H}_x^T \\ \tilde{H}_x & -\tilde{H}_x\tilde{f} \end{bmatrix}_{t=0} \begin{bmatrix} \delta x(0) \\ d\tau \end{bmatrix} \\ & + \delta\lambda^T(0)\delta x(0) - \delta\lambda^T(\tau)\delta x(\tau) \Bigg\}. \end{aligned} \tag{49}$$

To relate $\delta\lambda(0)$ and $\delta\lambda(\tau)$ to the given perturbations $\delta x(0)$ and $d\tau$, (19) is solved by using the state transition matrix:

$$\delta x(t) = \phi_{11}(t, 0)\delta x(0) + \phi_{12}(t, 0)\delta\lambda(0), \tag{50}$$

$$\delta\lambda(t) = \phi_{21}(t, 0)\delta x(0) + \phi_{22}(t, 0)\delta\lambda(0). \tag{51}$$

By substituting the boundary condition (14) into (50), where (50) is evaluated at τ, the following can be obtained:

$$\delta x(0) - \tilde{f}(\tau)\, d\tau = \phi_{11}(\tau, 0)\delta x(0) + \phi_{12}(\tau, 0)\delta\lambda(0). \tag{52}$$

Since $\phi_{12}(\tau, 0)$ is invertible, $\delta\lambda(0)$ can be determined as

$$\delta\lambda(0) = \phi_{12}^{-1}(\tau, 0)[I - \phi_{11}(\tau, 0)]\delta x(0) - \phi_{12}^{-1}(\tau, 0)\tilde{f}(\tau)\, d\tau. \tag{53}$$

Furthermore, by using (51) and (53), $\delta\lambda(\tau)$ is obtained in terms of $\delta x(0)$ and $d\tau$,

$$\begin{aligned} \delta\lambda(\tau) = & -\phi_{22}(\tau, 0)\phi_{12}^{-1}(\tau, 0)\tilde{f}(\tau)\, d\tau \\ & + \{\phi_{21}(\tau, 0) + \phi_{22}(\tau, 0)\phi_{12}^{-1}(\tau, 0)[I - \phi_{11}(\tau, 0)]\}\delta x(0). \end{aligned} \tag{54}$$

If (14), (53), and (54) are used in (49), the second variational cost $d^2\bar{J}^*(\delta x(0), d\tau)$ reduces to

$$d^2\bar{J}^*(\delta x(0), d\tau) = \frac{1}{\tau}[\delta x^T(0)\ d\tau]\bar{M}(0,\tau)\begin{bmatrix}\delta x(0)\\ d\tau\end{bmatrix}. \tag{55}$$

From (55), it follows that a necessary condition for the periodic path to be a minimum is $\bar{M}(0,\tau)\geqq 0$.

Remark. The necessary condition in Proposition 4 applies for all t_0. By replacing $t_0=0$ by any $t_0\in[0,\tau)$, $\bar{M}(t_0, t_0+\tau)$ can be defined the same as (48) except that ϕ_{ij} corresponds to $\phi_{ij}(t_0+\tau, t_0)$, and $\tilde{H}_x$ and $\tilde{f}$ are evaluated at t_0.

Perturbations along the periodic path do not change the orbit, only the part of perturbations perpendicular to the path contributes to a new path. By defining the projection operator $V(t)$ as

$$V = I - \tilde{f}[\tilde{f}^T\tilde{f}]^{-1}\tilde{f}^T, \tag{56}$$

the definition of strong positivity given in [13] is that there exist a positive number k and a positive definite matrix M such that

$$d^2\bar{J} \geqq k\int_0^\tau \delta x^T(t)V^T(t)MV(t)\delta x(t)\,dt. \tag{57}$$

This means that the change in cost from path $\tilde{x}(t)$ to $\tilde{x}(t)+\delta x(t)$ is dominated by the second variational cost for sufficiently small $\|\delta x\|$ [7].

PROPOSITION 5. *If $\xi^T\bar{M}(t, t+\tau)\xi\geqq 0$, for every $n+1$-vector ξ, and the equality is true only if $\xi^T=[\varepsilon\tilde{f}^T(t)\quad 0]$, where ε is an arbitrary real number, then the second variation is strongly positive.*

Proof. Define $\delta y(t)$ to be the component of $\delta x(t)$ in the space orthogonal to the velocity direction of the nominal periodic path,

$$\delta y(t) = V(t)\delta x(t). \tag{58}$$

Let $\sigma_i(t)$, $i=1, 2, \cdots, n+1$ be eigenvalues of $\bar{M}(t, t+\tau)$ such that $\sigma_1(t)=0$ and $\sigma_i(t)>0$, $i=2, \cdots, n+1$. Denote $\underline{\sigma}=\min_{i>1, 0\leqq t\leqq\tau}\sigma_i(t)$. Let L be an $(n+1)\times(n+1)$ orthonormal matrix consisting of the eigenvectors of $\bar{M}(t, t+\tau)$ such that

$$L^T\bar{M}L = \operatorname{diag}[0, \sigma_2, \cdots, \sigma_{n+1}]. \tag{59}$$

Then, L can be written in a partitioned form as

$$L = \begin{bmatrix}\varepsilon\tilde{f} & \bar{V}\\ 0 & \bar{W}\end{bmatrix} \tag{60}$$

where $\bar{V}$ and $\bar{W}$ are $n\times n$ matrix and $1\times n$ matrix, respectively, and ε is a real number which normalizes the velocity vector $\tilde{f}(t)$.

$$\begin{aligned} d^2\bar{J}^*(\delta x(t), d\tau) &= \frac{1}{\tau}[\delta x^T(t)\quad d\tau]\bar{M}(t, t+\tau)\begin{bmatrix}\delta x(t)\\ d\tau\end{bmatrix}\\ &= \frac{1}{\tau}[\delta x^T\quad d\tau]\begin{bmatrix}\bar{V}\\ \bar{W}\end{bmatrix}\operatorname{diag}[\sigma_2, \cdots, \sigma_{n+1}][\bar{V}^T\quad \bar{W}^T]\begin{bmatrix}\delta x\\ d\tau\end{bmatrix}\\ &\geqq \frac{\underline{\sigma}}{\tau}\|\bar{V}^T\delta x + \bar{W}^T d\tau\|^2. \end{aligned} \tag{61}$$

Since L is orthonormal, the column vectors of L are orthogonal. By examining (60), the orthogonality implies that $\tilde{f}$ is orthogonal to the column vectors of $\bar{V}$ and $\bar{W}^T$ is orthogonal to the column vectors of $\bar{V}^T$. Therefore,

$$\bar{V}^T V = \bar{V}^T[I - \tilde{f}(\tilde{f}^T\tilde{f})^{-1}\tilde{f}] = \bar{V}^T, \tag{62}$$

and

$$\|\bar{V}^T\delta x + \bar{W}^T d\tau\|^2 = \|\bar{V}^T\delta x\|^2 + \|\bar{W}^T d\tau\|^2 \geqq \|\bar{V}^T\delta x\|^2. \tag{63}$$

By using (61), (62), and (63), $d^2\bar{J}^*(\delta x(t), d\tau)$ is related to δy as follows:

$$d^2\bar{J}^*(\delta x(t), d\tau) \geqq \frac{\sigma}{\tau}\|\bar{V}^T V\delta x\|^2 = \frac{\sigma}{\tau}\delta y^T\bar{V}\bar{V}^T\delta y. \tag{64}$$

From the orthogonality, $\tilde{f}$ is an eigenvector of $\bar{V}\bar{V}^T$ corresponding to a zero eigenvalue. Furthermore, rank $\bar{V}\bar{V}^T = n-1$. Otherwise, L defined in (60) is not invertible. Let $\mu_1(t), \cdots, \mu_n(t)$ be eigenvalues of $\bar{V}\bar{V}^T$ such that $\mu_1 = 0$ and $\mu_i > 0$, $i = 2, \cdots, n$. Denote $\underline{\mu} = \min_{i>1, t\in[0,\tau]}\mu_i(t)$. Since δy is orthogonal to $\tilde{f}$, then

$$\frac{\sigma}{\tau}\delta y^T\bar{V}\bar{V}^T\delta y \geqq \frac{\sigma\underline{\mu}}{\tau}\|\delta y\|^2. \tag{65}$$

Let $M = I$, $k = (\sigma\underline{\mu})/\tau^2$, then by the mean value theorem, there exists $\theta \in (0, \tau)$ such that

$$k\int_0^\tau \delta x^T(t)V^T(t)MV(t)\delta x(t)\,dt = \frac{\sigma\underline{\mu}}{\tau^2}\int_0^\tau \|\delta y(t)\|^2\,dt = \frac{\sigma\underline{\mu}}{\tau}\|\delta y(\theta)\|^2. \tag{66}$$

Let $\delta x(\theta)$ be the state variation of the perturbed path at $t = \theta$. Then $d^2\bar{J}^*(\delta x(\theta), d\tau)$ is the accessory minimum corresponding to $t_0 = \theta$ with initial perturbation $\delta x(\theta)$ and $d\tau$. Since the neighboring optimal control is used for constructing a closed orbit starting at $\delta x(\theta)$, $d^2\bar{J}^*(\delta x(\theta), d\tau)$ is less than or equal to the cost of an arbitrary perturbed path passing that point. Therefore, by combining (64) with (66),

$$\frac{\sigma\underline{\mu}}{\tau}\|\delta y(\theta)\|^2 \leqq d^2\bar{J}^*(\delta x(\theta), d\tau) \leqq d^2\bar{J}. \tag{67}$$

Remark. The restrictions in [13] that the eigenvalues of the monodromy matrix be distinct except for two unit eigenvalues and the existence of periodic Riccati solution, which are required for the proof of the strongly positive condition, are no longer needed here for a single-period process.

PROPOSITION 6. *If the conditions of Proposition 5 hold, then the two unit eigenvalues of the monodromy matrix are coupled in the same Jordan box.*

Proof. Suppose that $\begin{bmatrix}\alpha\\ \gamma\end{bmatrix}$ is a primary eigenvector of the monodromy matrix corresponding to the unit eigenvalue. By using the symplectic property of $\phi(t, 0)$, the following can be obtained:

$$\alpha^T[\phi_{12}^{-1}(I - \phi_{11}) + \phi_{12}^{-T}(I - \phi_{22}^T)]\alpha = 0. \tag{68}$$

Furthermore, if $\begin{bmatrix}\alpha\\ \gamma\end{bmatrix}$ is an eigenvector which is independent of the one along the velocity direction, i.e.,

$$\begin{bmatrix}\alpha\\ \gamma\end{bmatrix} \neq \varepsilon\begin{bmatrix}\tilde{f}\\ -\tilde{H}_x^T\end{bmatrix},$$

where ε is any real number, then two possible cases may occur:

(i) $\alpha \neq \varepsilon f$. In this case, the strongly positiveness does not hold;

(ii) $\alpha = \varepsilon \tilde{f}$ for some ε, but $\gamma \neq -\varepsilon \tilde{H}_x^T$. In this case, a contradiction to the invertibility of $\phi_{12}(\tau, 0)$ is produced. Since

$$\begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} - \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} \varepsilon \tilde{f} \\ -\varepsilon \tilde{H}_x^T \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} 0 \\ \gamma + \varepsilon \tilde{H}_x^T \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma + \varepsilon \tilde{H}_x^T \end{bmatrix}, \tag{69}$$

this indicates that $\phi_{12}(\gamma + \varepsilon \tilde{H}_x^T) = 0$. The result contradicts the invertibility assumption on ϕ_{12}.

Therefore, the unit eigenvalues of the monodromy matrix must be coupled in the same Jordan box. □

The importance of this section has been to establish a second-order necessary condition for the weak local optimality of a periodic process. The condition is derived in a simple algebraic form. From this algebraic condition some properties of the periodic process are determined under weaker assumptions than used in [13].

6. Summary of necessary and sufficient conditions. Based on the discussions in §§ 4 and 5, a set of second variational necessary and sufficient conditions for local optimality of a single-period process are stated.

Sufficient Condition. A second-order sufficient condition for a periodic path to be a weak local minimum given Assumptions 1–6 is that:

(i) For all $t_0 \in [0, \tau)$, there exists a continuous real symmetric solution to the Riccati differential equation (47) on $t \in [t_0, t_0 + \tau]$;

(ii) For all $t_0 \in [0, \tau)$ and $n+1$-vectors ξ, $\xi^T \bar{M}(t_0, t_0 + \tau)\xi \geqq 0$, and the equality is true only if $\xi = [\varepsilon \tilde{f}^T(t_0)\ 0]^T$, where ε is a real number.

Necessary Condition. A second-order necessary condition for a periodic path to be a weak local minimum given Assumptions 1–6 is the same as the sufficient condition, except that condition (ii) is weakened as: For all $t_0 \in [0, \tau)$, $\bar{M}(t_0, t_0 + \tau) \geqq 0$.

The necessary and sufficient conditions given above apply for weak local optimality of a τ-period process called a single periodic process. Condition (i) is a conjugate point condition associated with optimality of the control $u(\cdot)$. Condition (ii) ensures the nonnegativity of the second variation when the period and the initial state are perturbed. The invertibility of ϕ_{12} is insured by condition (i). Note that there is no requirement that the solution of the Riccati equation be periodic as there is in [3], [6], [13]. For the sufficient conditions, the requirement on the eigenvalues of the monodromy matrix is that there are no uncoupled unit eigenvalues, therefore the conditions in [13] are weakened.

7. Second variation analysis for an infinitely-repeated periodic process. In §§ 3–6, the second variation conditions for local optimality of a single periodic process are considered. If a periodic path is an extremal for the optimal periodic control problem with period τ, then for any positive integer k, the $k\tau$-periodic path, which is obtained by repeating the orbit k times, is also an extremal for the same problem, since the necessary conditions of Proposition 1 are satisfied. The optimal control problem for a k-repeated periodic process is defined to minimize the performance criterion

$$J(u(\cdot), x(0), \tau) = \frac{1}{k\tau} \int_0^{k\tau} L(x(t), u(t))\, dt \tag{70}$$

with respect to $u(\cdot)$, $x(0)$, and τ, subject to the dynamic equation (2), and the periodic boundary conditions

$$x(i\tau) = x(0), \quad i = 1, 2, \cdots, k. \tag{71}$$

In the first-order variation analysis, there is no basic difference between the problem described above and the problem stated in (1) to (3). But when small variations around these repeated periodic paths are considered, some interesting phenomena can be revealed. For example, now a comparison path may be closed for the first time at $t = k\tau$ instead of closed around $t = \tau$. In this section, small variations with respect to the k-repeated periodic path are considered. If these repeated periodic processes are included, the second variation conditions summarized in § 6 must be strengthened. In particular, the existence of a periodic solution to the Riccati differential equation is required if k becomes infinite.

If the time interval $(t_0, t_0+\tau]$ is replaced by the time interval $(t_0, t_0+k\tau]$, the conjugate point conditions proved in § 4 apply without modification to each k-repeated τ-periodic process. That is, if $\phi_{12}(t_c, t_0)$ is not invertible for $t_c \in (t_0, t_0+k\tau]$ for some $k>1$, a $k\tau$-periodic path conjugate to the k-repeated τ-periodic extremal path can be found having the same cost as the periodic extremal path. Then, a nonextremal path with period $k\tau$ can be found as shown in § 4, and the cost of this nonextremal path is less than that of the conjugate path. Therefore, the k-repeated τ-periodic path is not a minimum.

The absence of a point $t_c \in (t_0, t_0+k\tau]$ conjugate to t_0 for each t_0 is a necessary condition for the optimality of a k-repeated periodic process. By letting k go to infinity, Proposition 3 and Corollary 1 are modified for the infinitely-repeated periodic process to the following necessary condition.

PROPOSITION 7. *Given Assumptions* 1–4, *a necessary condition for optimality of an infinitely-repeated periodic process is that there exists a continuous real symmetric solution to the Riccati differential equation* (47) *for all t.*

Next, the necessity of existence of a periodic Riccati solution for optimality is shown.

PROPOSITION 8. *Given that* $\phi_{12}(t, t_0)$ *is invertible for* $t > t_0$, *a necessary condition for optimality of an infinitely-repeated periodic process is that the matrix* $-\phi_{12}^{-1}(t, t_0)\phi_{11}(t, t_0)$ *is bounded from below for* $t > t_0$.

Proof. Let the initial state perturbation at t_0 be $\delta x(t_0) = \delta x_0$ and solve the following accessory minimum problem: Minimize the performance criterion

$$J_2(t_0, t_f) = \int_{t_0}^{t_f} [\delta x^T \quad \delta u^T] \begin{bmatrix} \tilde{H}_{xx} & \tilde{H}_{xu} \\ \tilde{H}_{ux} & \tilde{H}_{uu} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta u \end{bmatrix} dt \tag{72}$$

subject to

$$\delta \dot{x} = \tilde{f}_x \delta x + \tilde{f}_u \delta u \tag{73}$$

and

$$\delta x(t_0) = \delta x_0, \quad \delta x(t_f) = 0. \tag{74}$$

According to (28), the cost for this accessory minimum problem is

$$J_2(t_0, t_f) = \delta x^T(t_0) \delta \lambda(t_0). \tag{75}$$

Since

$$\delta x(t_f) = 0 = \phi_{11}(t_f, t_0)\delta x_0 + \phi_{12}(t_f, t_0)\delta \lambda(t_0), \tag{76}$$

$\delta \lambda(t_0)$ can be written as

$$\delta \lambda(t_0) = -\phi_{12}^{-1}(t_f, t_0)\phi_{11}(t_f, t_0)\delta x_0. \tag{77}$$

By substituting (77) into (75),

$$J_2(t_0, t_f) = -\delta x_0^T \phi_{12}^{-1}(t_f, t_0)\phi_{11}(t_f, t_0)\delta x_0. \tag{78}$$

Since the system is controllable, given $t_1 < t_0$, a finite control can be found to take the variation from $\delta x(t_1) = 0$ to $\delta x(t_0) = \delta x_0$, and the corresponding second variational cost is bounded,

$$J_2(t_1, t_0) = \delta x_0^T D \delta x_0 \tag{79}$$

where D is a bounded matrix. If $-\phi_{12}^{-1}(t_f, t_0)\phi_{11}(t_f, t_0)$ is not bounded from below, there exist some δx_0 and t_f in the interval $t_1 < t_f < t_1 + K\tau$ for some K, such that

$$J_2(t_1, t_f) = J_2(t_1, t_0) + J_2(t_0, t_f) < 0, \tag{80}$$

i.e., there exists a $K\tau$-periodic path whose cost is less than that of the K-repeated τ-periodic extremal path. Therefore, the K-repeated τ-periodic path is not a minimum. Clearly this holds for all $k > K$, in particular, as k goes to infinity. □

Note that t_0 represents a particular point on the path which is held fixed. As t_f increases, the cost (78) is shown in the next proposition to monotonically decrease since the terminal state variations are constrained to zero as given in (74).

PROPOSITION 9. *A necessary condition for an infinitely-repeated periodic path to be a minimum is that there exists a continuous real symmetric periodic solution to the Riccati differential equation* (47).

Proof. By differentiating $-\phi_{12}^{-1}(\theta, t)\phi_{11}(\theta, t)$ with respect to θ and manipulating terms using the symplectic property of transition matrix [12], the following matrix differential equation can be obtained:

$$\frac{d}{d\theta}[-\phi_{12}^{-1}(\theta, t)\phi_{11}(\theta, t)] = -\phi_{12}^{-1}(\theta, t)B(\theta)\phi_{12}^{-T}(\theta, t). \tag{81}$$

Since $B(\theta) \geqq 0$, (81) shows that for every t, $-\phi_{12}^{-1}(\theta, t)\phi_{11}(\theta, t)$ is a nonincreasing matrix with respect to θ. According to Proposition 8, $-\phi_{12}^{-1}(\theta, t)\phi_{11}(\theta, t)$ is bounded from below. Therefore, $-\phi_{12}^{-1}(\theta, t)\phi_{11}(\theta, t)$ approaches a limit as θ approaches positive infinity. Denote this limit by $\underline{P}(t)$, then

$$\underline{P}(t) = \lim_{\theta \to \infty} -\phi_{12}^{-1}(\theta, t)\phi_{11}(\theta, t). \tag{82}$$

Since ϕ is the transition matrix of a periodic Hamiltonian system of differential equations, then

$$\phi(\theta, t + \tau) = \phi(\theta - \tau, t) \tag{83}$$

and

$$\underline{P}(t + \tau) = \lim_{\theta \to \infty} -\phi_{12}^{-1}(\theta - \tau, t)\phi_{11}(\theta - \tau, t) = \underline{P}(t). \tag{84}$$

Therefore $\underline{P}(t)$ is periodic.

Finally, $\underline{P}(t)$ satisfying (47) needs to be shown. By differentiating $-\phi_{12}^{-1}(\theta, t)\phi_{11}(\theta, t)$ with respect to t, it is shown that for every θ, $-\phi_{12}^{-1}(\theta, t)\phi_{11}(\theta, t)$ satisfies (47) [12], where $t < \theta$ is assumed. By letting θ go to infinity, it is concluded that $\underline{P}(t)$ is a periodic solution to (47).

PROPOSITION 10. *The controllability assumption and the existence of a continuous real symmetric periodic solution to the Riccati differential equation imply the nonnegativeness of the second variation, or equivalently, imply* $\bar{M}(t_0, t_0 + k\tau) \geqq 0$ *for* $k = 1, 2, \cdots$.

Proof. Only the case $t_0 = 0$ is proved, since the procedure is the same for other t_0. By adding the zero term

$$0 = \frac{1}{k\tau}\int_0^{k\tau} \delta x^T \underline{P}(\tilde{f}_x \delta x + \tilde{f}_u \delta u - \delta \dot{x})\, dt \tag{85}$$

to (12) with τ and $d\tau$ replaced by $k\tau$ and $d(k\tau)$, and following the mathematical manipulation, the second variational cost can be written as

$$d^2\bar{J}=\frac{1}{k\tau}\left\{[\delta x^T(0)\quad d(k\tau)]\begin{bmatrix}0 & (\tilde{H}_x+\tilde{f}^T\underline{P})^T\\ \tilde{H}_x+\tilde{f}^T\underline{P} & (\tilde{H}_x+\tilde{f}^T\underline{P})\tilde{f}\end{bmatrix}_{t=0}\begin{bmatrix}\delta x(0)\\ d(k\tau)\end{bmatrix}\right.$$
$$\left.+\int_0^{k\tau}\|\tilde{H}_{uu}^{-1}(\tilde{f}_u^T\underline{P}+\tilde{H}_{ux})\delta x+\delta u\|_{H_{uu}}^2\,dt\right\}. \tag{86}$$

Consider the following similarity transformation

$$\begin{bmatrix}I & 0\\ -\underline{P}(0) & I\end{bmatrix}\begin{bmatrix}\phi_{11} & \phi_{12}\\ \phi_{21} & \phi_{22}\end{bmatrix}_{t=k\tau}\begin{bmatrix}I & 0\\ \underline{P}(0) & I\end{bmatrix}=\begin{bmatrix}\phi_{11}+\phi_{12}\underline{P} & \phi_{12}\\ 0 & -\underline{P}\phi_{12}+\phi_{22}\end{bmatrix}_{t=k\tau}$$
$$\triangleq\begin{bmatrix}\phi_z & \phi_{12}\\ 0 & \phi_z^{-T}\end{bmatrix}_{t=k\tau} \tag{87}$$

where $Z=A-B\underline{P}$, ϕ_{ij}, $i,j=1,2$, are partitioned blocks of $\phi(k\tau,0)$, and ϕ_Z is the transition matrix associated with Z. Then,

$$\begin{bmatrix}\phi_{11} & \phi_{12}\\ \phi_{21} & \phi_{22}\end{bmatrix}\begin{bmatrix}\tilde{f}\\ -\tilde{H}_x^T\end{bmatrix}=\begin{bmatrix}\tilde{f}\\ -\tilde{H}_x^T\end{bmatrix} \tag{88}$$

implies

$$\phi_z\tilde{f}-\phi_{12}(\underline{P}\tilde{f}+\tilde{H}_x^T)=\tilde{f}. \tag{89}$$

As shown in [14], the controllability assumption and the existence of a real symmetric periodic solution to the Riccati differential equation imply that the monodromy matrix has no uncoupled unit eigenvalues. The eigenvector structure of $\underline{P}$ must include all the primary eigenvectors of the unit eigenvalues. Therefore, $\tilde{f}$ is an eigenvector of unit eigenvalue of the closed-loop matrix ϕ_z, i.e.,

$$\phi_z(k\tau,0)\tilde{f}(0)=\tilde{f}(0). \tag{90}$$

Since ϕ_{12} is invertible, from (89) and (90), the following is obtained:

$$\underline{P}\tilde{f}+\tilde{H}_x^T=0. \tag{91}$$

By using (91) in (86), the first term in the right-hand side of (86) then vanishes, therefore, $d^2\bar{J}\geqq 0$. □

PROPOSITION 11. *If there exists a continuous real symmetric periodic solution to the Riccati differential equation, and the eigenvalues of the monodromy matrix are off the unit circle except for one pair of unit eigenvalues, then the second variation is strongly positive.*

Proof. From (86), the second variation is zero if and only if

$$\delta u=-\tilde{H}_{uu}^{-1}(\tilde{f}_u^T\underline{P}+\tilde{H}_{ux})\delta x. \tag{92}$$

Then

$$\delta\dot{x}=\tilde{f}_x\delta x-\tilde{f}_u\tilde{H}_{uu}^{-1}(\tilde{f}_u^T\underline{P}+\tilde{H}_{ux})\delta x \tag{93}$$

and

$$\delta x(t)=\phi_z(t,0)\delta x(0). \tag{94}$$

By using the periodic boundary condition

$$\delta x(k\tau)=\delta x(0)-\tilde{f}(0)d(k\tau), \tag{95}$$

equation (94) evaluated at $t = k\tau$ can be written as

$$\delta x(0) - \tilde{f}(0)d(k\tau) = \phi_z(k\tau, 0)\delta x(0), \tag{96}$$

or

$$[\phi_z(k\tau, 0) - I\tilde{f}(0)]\begin{bmatrix}\delta x(0)\\ d(k\tau)\end{bmatrix} = 0. \tag{97}$$

It can be shown that $[\delta x^T(0)\ d(k\tau)]^T = [\varepsilon\tilde{f}^T(0)\ 0]^T$ satisfies (97), and it will be shown that it is the only solution. Since $\phi_z(k\tau, 0) = \phi_z^k(\tau, 0)$ has only one unit eigenvalue, therefore

$$\operatorname{rank}[\phi_z(k\tau, 0) - I] = n - 1. \tag{98}$$

Also, $\tilde{f}(0)$ belongs to the null-space of $(\phi_z(k\tau, 0) - I)$. Therefore

$$\operatorname{rank}[\phi_z(k\tau, 0) - I\tilde{f}(0)] = n. \tag{99}$$

This implies that $\bar{M}(0, k\tau)$ has strongly positive property, i.e., $\xi^T\bar{M}\xi = 0$ only if $\xi = [\varepsilon\tilde{f}^T\ 0]^T$. Therefore, the second variation is strongly positive. □

8. New necessary and sufficient conditions. The results of § 7 are summarized into a new set of necessary and sufficient conditions for local weak optimality of an infinitely-repeated periodic process.

PROPOSITION 12. *The necessary condition for local weak optimality of an infinitely-repeated periodic process given Assumptions* 1-6 *is that there exists a continuous real symmetric periodic solution to the Riccati differential equation* (47); *the sufficient condition to the same problem is that in addition, the monodromy matrix has no eigenvalues on the unit circle except for the pair of unit eigenvalues.*

The conditions given in Proposition 12 generalize and extend the previous results.

9. Conclusions. Second variational necessary and sufficient conditions for weak local optimality of a periodic process are discussed. The conditions are derived from two aspects: conjugate point condition and nonnegative condition. The conjugate point condition is related to the existence of a real symmetric solution to the Riccati differential equation over the period, and the nonnegative condition is given in the form of a matrix condition obtained from solving the accessory minimum problem. The weak form of the conditions imply necessity, and the strong form of the conditions imply sufficiency. Some other properties related to a periodic process are also derived under weaker assumptions, such as the strongly positive property (Proposition 5) and the coupling of the two unit eigenvalues of the monodromy matrix (Proposition 6). If the variations around an infinitely-repeated periodic path are considered, the existence of a real symmetric periodic solution to the Riccati differential equation is necessary for optimality and implies nonnegativity of the second variational cost (Propositions 7 to 10). The existence of a periodic solution to the Riccati differential equation and the requirement that no other eigenvalues of the monodromy matrix are on the unit circle except for a pair of unit eigenvalues are shown to be sufficient for optimality (Proposition 11). A new set of necessary and sufficient conditions is given (Proposition 12) which is compact and convenient to use.

Acknowledgment. The authors are grateful to Professor Steven I. Marcus for his helpful comments on this paper.

REFERENCES

[1] D. S. BERNSTEIN AND E. G. GILBERT, *Optimal periodic control: The π test revisited*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 673-684.

[2] S. BITTANTI, G. FRONZA, AND G. GUARDABASSI, *Periodic control: A frequency domain approach*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 33-38.
[3] S. BITTANTI, A. LOCATELLI, AND C. MAFFEZZONI, *A second variation method in periodic optimization*, J. Optim. Theory Appl., 14 (1974), pp. 31-39.
[4] J. V. BREAKWELL AND Y. C. HO, *On the conjugate point condition for the control problem*, Internat. J. Engrg. Sci., 2 (1965), pp. 565-579.
[5] A. E. BRYSON AND Y. C. HO, *Applied Optimal Control*, Hemisphere Publishing, New York, 1975.
[6] K. S. CHANG, *Second variation for periodic optimization problems*, in Periodic Optimization, Vol. II, CISM Lectures 135, A. Marzollo, ed., Springer-Verlag, New York, 1972.
[7] I. M. GELFAND AND S. V. FOMIN, *Calculus of Variation*, Prentice Hall, Englewood Cliffs, NJ, 1963.
[8] E. G. GILBERT, *Optimal periodic control: a general theory of necessary conditions*, SIAM J. Control Optim., 15 (1977), pp. 717-746.
[9] F. J. M. HORN AND R. C. LIN, *Periodic processes: a variational approach*, Indust. Eng. Chem. Proc. Design. Dev., 6 (1967), pp. 21-30.
[10] W. T. REID, *Riccati Differential Equations*, Academic Press, New York, 1972.
[11] J. L. SPEYER, *Nonoptimality of the steady-state cruise for aircraft*, AIAA J., 14 (1976), pp. 1604-1610.
[12] ———, *The linear-quadratic control problem*, in Control and Dynamic Systems, Academic Press, New York, 1986, pp. 241-293.
[13] J. L. SPEYER AND R. T. EVANS, *A second variational theory for optimal periodic processes*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 138-148.
[14] Q. WANG AND J. L. SPEYER, *The periodic Riccati differential equation and the periodic regulator*, Proc. IEEE Conf. Decision and Control, 1987, pp. 288-292.
[15] V. ZEIDAN AND P. ZEZZA, *Necessary conditions for optimal control problems: conjugate points*, SIAM J. Control Optim., 26 (1988), pp. 592-608.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 2, pp. 498-501, March 1990

015

EXACT CONTROLLABILITY: COEFFICIENT DEPENDING ON THE TIME*

JAIME E. MUÑOZ RIVERA

Abstract. The following is considered: The wave equation $y''-a(t)\Delta y=0$, with initial data given by (y_0, y_1) in $L^2(\Omega)\times H^{-1}(\Omega)$, and the nonhomogeneous condition $y=v$ on the boundary Σ of $Q=\Omega\times]0, T[$. Exact controllability means that there exist a time $T'>0$, and a control v such that $y(T', v)=y'(T', v)=0$. The main result of this paper is to prove that the above system is exactly controllable when $a(\cdot)$ is a monotonic function in some interval of width greater than $R\sqrt{\|a\|_\infty}/a_0$, where $\|a\|_\infty=\sup\{|a(t)|;\ t\in\mathbb{R}\}$, $a_0\leqq a(t)$ for all $t\in\mathbb{R}$ and $R=\sup\{\|x-x_0\|;\ x\in\Omega\}$ for some fixed x_0 in $\mathbb{R}^n$.

Key words. exact controllability, control of distributed systems, wave equation, linear equation

1. Introduction. Let Ω be an open bounded set of $\mathbb{R}^n$ with boundary Γ of class C^2. Let Q be a cylinder defined by $Q=\Omega\times]0, T[$ and Σ be the lateral boundary of Q given by $\Sigma=\Gamma\times]0, T[$. We denote by $(\cdot,\cdot)_\Omega$ and $\|\cdot\|_\Omega$ the inner product and the norm of $L^2(\Omega)$, respectively. In [2], the author has considered the exact controllability problem for the system

$$\begin{aligned} &y''-a(t)\Delta y=0 && \text{in } Q,\\ &y(x,t)=v && \text{on } \Sigma,\\ &y(0)=y_0,\quad y'(0)=y_1 && \text{in } \Omega \end{aligned} \tag{1.1}$$

where a is a function satisfying

$$a, a'\in L^\infty(\mathbb{R}_+),\qquad a(t)\geqq a_0>0. \tag{1.2}$$

We want to find a control v of class $L^2(\Sigma)$, driving the system to rest; that is, we want to find an element v in $L^2(\Sigma)$ satisfying

$$y(v;T)=y'(v,T)=0\quad \text{for } T>0. \tag{1.3}$$

When this is the case, we say that the system is exactly controllable. In [2] it is shown that (1.3) is valid when $a'(t)\geqq 0$ for all $t\geqq 0$. The main result of this paper is to prove that the system (1.1) is exactly controllable when a is a monotonic function in some interval $[T_0, T_1]$ such that

$$T_1-T_0>\frac{R\sqrt{\|a\|_\infty}}{a_0}. \tag{1.4}$$

2. The main result. Let us define ϕ as the solution of the following system:

$$\begin{aligned} &\phi''-a(t)\Delta\phi=0 && \text{in } Q,\\ &\phi(x,t)=0 && \text{on } \Sigma,\\ &\phi(0)=\phi_0,\quad \phi'(0)=\phi_1 && \text{in } \Omega. \end{aligned} \tag{2.1}$$

First, we will prove that there exists a positive constant C satisfying the following inequality:

$$CE(0)\leqq\frac{1}{2}R\|a\|_\infty\int_{\Sigma_0}\left|\frac{\partial\phi}{\partial\nu}\right|^2 d\Sigma. \tag{2.2}$$

* Received by the editors December 19, 1988; accepted for publication May 26, 1989.

† National Laboratory of Scientific Computation, Rua Lauro Muller 455, Botafogo 22290, Rio de Janeiro, Brasil.

Here $R=\sup\{\|x-x_0\|;\, x\in\Omega\}$, where x_0 is a fixed point of $\mathbb{R}^n$, and $\Sigma_0=\Gamma_0\times]0,\, T[$ where $\Gamma_0=\{x\in\Gamma:\,(x-x_0)\cdot\nu(x)\geqq 0\}$. The following remarks are in order.

Remark 2.1. Put $E(t)=\frac{1}{2}\{\|\phi'(t)\|_\Omega^2+a(t)\|\nabla\phi(t)\|_\Omega^2\}$. From (2.1) it follows that

$$E'(t)=\tfrac{1}{2}a'(t)\|\nabla\phi\|_\Omega^2\geqq-\tfrac{1}{2}|a'(t)|\,\|\nabla\phi(t)\|_\Omega^2.$$

From this inequality and by (1.2), we obtain

$$E'(t)\geqq-\frac{|a'(t)|}{2a_0}E(t). \tag{2.3}$$

Remark 2.2. For all $t\in[0,\, T_0]$ we have that

$$E(t)\geqq\exp\left\{-\frac{1}{2a_0}\int_0^{T_0}|a'(s)|\,ds\right\}E(0).$$

In fact, if we multiply (2.3) by $\rho(t)=\exp\{1/2a_0\int_0^t|a'(s)|\,ds\}$, then we have $d/dt\,\{E(t)\rho(t)\}\geqq 0$. Integrating from zero to t, we obtain the result.

Remark 2.3. If a is a monotonic function on $[T_0,\, T_1]$, then we have

$$\sqrt{a_0/\|a\|_\infty}\,E(T_0)\leqq E(t)\leqq E(T_0)\quad\forall t\in[T_0,\, T_1],\text{ when } a'(t)\leqq 0, \tag{2.4}$$

$$E(T_0)\leqq E(t)\leqq\sqrt{\|a\|_\infty/a_0}\,E(T_0)\quad\forall t\in[T_0,\, T_1],\text{ when } a'(t)\geqq 0. \tag{2.5}$$

In fact, let us suppose that a is a nondecreasing function; then we have

$$0\leqq E'(t)=\frac{1}{2}a'(t)\|\nabla\phi(t)\|_\Omega^2\leqq\frac{1}{2}\frac{a'(t)}{a(t)}E(t).$$

From Gronwall's inequality we obtain

$$E(T_0)\leqq E(t)\leqq\sqrt{a(T_1)/a(T_0)}\,E(T_0)\leqq\sqrt{\|a\|_\infty/a_0}\,E(T_0)$$

from which (2.5) follows. Using the same reasoning, we can prove that relation (2.4) is also valid.

Now we are able to prove the following lemma.

LEMMA 2.1. *Let a be a monotonic function on $[T_0,\, T_1]$ satisfying* (1.2). *Then there exists a constant C satisfying condition* (2.2).

Proof. Set $Q'=\Omega\times]T_0,\, T[$, $\Sigma'=\Gamma\times]T_0,\, T[$ and put $m(x)=x-x_0$. Multiplying system (2.1) by $m_k(\partial\phi/\partial x_k)$, we have

$$\begin{aligned}\frac{1}{2}\int_{\Sigma'}a(t)\left|\frac{\partial\phi}{\partial\nu}\right|^2 m_k\nu_k\,d\Sigma&=\left[\left(\phi'(t),\, m_k\frac{\partial\phi}{\partial x_k}(t)\right)_\Omega\right]_{T_0}^{T}-\frac{n}{2}\int_{Q'}\{|\phi'|^2-a(t)|\nabla\phi|^2\}\,dx\,dt\\&\quad+\int_{Q'}a(t)|\nabla\phi|^2\,dx\,dt\end{aligned}$$

where $m(x)=(m_1(x),\cdots,m_k(x),\cdots,m_n(x))$.

But

$$\int_{Q'}\{|\phi'(t)|^2-a(t)|\nabla\phi(t)|^2\}\,dx\,dt=[(\phi'(t),\,\phi(t))_\Omega]_{T_0}^{T}$$

from which we obtain that

$$\frac{1}{2}\int_{\Sigma'} a(t)\left|\frac{\partial\phi}{\partial\nu}\right|^2 m_k\nu_k\,d\Sigma=\left[\left(\phi'(t), m_k\frac{\partial\phi}{\partial x_k}(t)+\frac{n-1}{2}\phi(t)\right)_\Omega\right]_{T_0}^{T}+\int_{T_0}^{T}E(t)\,dt. \tag{2.6}$$

Defining $Z=(\phi'(t), m_k(\partial\phi/\partial x_k)(t)((n-1)/2)\phi(t))_\Omega$, we have

$$|Z|\leqq\|\phi'(t)\|_\Omega\left\|m_k\frac{\partial\phi}{\partial x_k}(t)+\frac{n-1}{2}\phi(t)\right\|_\Omega. \tag{2.7}$$

On the other hand

$$\left\|m_k\frac{\partial\phi}{\partial x_k}(t)+\frac{n-1}{2}\phi(t)\right\|_\Omega^2=\left\|m_k\frac{\partial\phi}{\partial x_k}(t)\right\|_\Omega^2+(n-1)\left(\phi(t), m_k\frac{\partial\phi}{\partial x_k}(t)\right)_\Omega+\frac{(n-1)^2}{4}\|\phi(t)\|_\Omega^2.$$

But

$$\left(\phi(t), m_k\frac{\partial\phi}{\partial x_k}(t)\right)_\Omega=-\frac{n}{2}\|\phi(t)\|_\Omega^2$$

from which we conclude

$$\left\|m_k\frac{\partial\phi}{\partial x_k}(t)+\frac{n-1}{2}\phi(t)\right\|_\Omega\leqq R\|\nabla\phi(t)\|_\Omega. \tag{2.8}$$

By (2.7) and (2.8) we have

$$[|Z|]_{T_0}^{T}\leqq\frac{R}{\sqrt{a_0}}\cdot\sup_{t\in[T_0,T]}E(t). \tag{2.9}$$

Set $\Sigma_0'=\Gamma_0\times]T_0, T[$, and suppose that $a'(t)\leqq 0$ on $[T_0, T_1]$. Now from (2.6), (2.7), (2.9), and Remark 2.3 we have

$$\begin{aligned}\frac{1}{2}R\|a\|_\infty\int_{\Sigma_0}\left|\frac{\partial\phi}{\partial\nu}\right|^2 d\Sigma&\geqq\frac{1}{2}R\|a\|_\infty\int_{\Sigma_0'}\left|\frac{\partial\phi}{\partial\nu}\right|^2 d\Sigma\geqq\frac{1}{2}\int_{\Sigma'}m_k\nu_k\left|\frac{\partial\phi}{\partial\nu}\right|^2 d\Sigma\\&\geqq-\frac{R}{\sqrt{a_0}}E(T_0)+\int_{T_0}^{T}E(t)\,dt\\&\geqq\left\{-\frac{R}{\sqrt{a_0}}+(T_1-T_0)\sqrt{\frac{a_0}{\|a\|_\infty}}\right\}E(T_0)+\int_{T_1}^{T}E(t)\,dt\\&\geqq\frac{1}{\sqrt{a_0}}\left\{-R+(T_1-T_0)\frac{a_0}{\sqrt{\|a\|_\infty}}\right\}E(T_0)\quad\text{for } T\geqq T_1.\end{aligned}$$

Finally the result follows from Remark 2.2. With the same reasoning we can argue the case for $a'(t)\geqq 0$, and the proof is thus complete. □

THEOREM 2.1. *Let $a(\cdot)$ as in Lemma 2.1. Then for all $\{y_0, y_1\}$ in $L^2(\Omega)\times H^{-1}(\Omega)$, there exist v in $L^2(\Sigma)$ such that y, the solution of system (1.1), satisfies (1.3) for $T>T_1$.*

Proof. Let us define ψ as the solution of

$$\begin{aligned}&\psi''-a(t)\Delta\psi=0 && \text{in } Q,\\&\psi(x,t)=w(x,t) && \text{on } \Sigma,\\&\psi(T)=\psi'(T)=0 && \text{in } \Omega,\end{aligned} \tag{2.10}$$

where $w=\partial\phi/\partial\nu$ on Σ_0, and $w=0$ on $\Sigma\backslash\Sigma_0$. It is well known that $\partial\phi/\partial\nu\in L^2(\Sigma)$ when $\{\phi_0, \phi_1\}\in H_0^1(\Omega)\times L^2(\Omega)$ and system (2.10) has only one solution, $\psi\in C([0, T]; L^2(\Omega))$.

Multiplying system (2.1) by ψ and applying Green's formulas, we have

$$\int_{\Sigma_0} a(t)\left|\frac{\partial\phi}{\partial\nu}\right|^2 d\Sigma = \langle\psi'(0), \phi_0\rangle - \langle\psi(0), \phi_1\rangle. \tag{2.11}$$

If we define the operator Λ as

$$\Lambda\{\phi_0, \phi_1\} = \{\psi'(0), -\psi(0)\} \tag{2.12}$$

for $\{\phi_0, \phi_1\}$ in $C_0^\infty(\Omega)\times C_0^\infty(\Omega)$, from Lemma 2.1 we conclude that the functional

$$\{\phi_0, \phi_1\} \mapsto \langle\Lambda\{\phi_0, \phi_1\}, \{\phi_0, \phi_1\}\rangle \tag{2.13}$$

defines a norm in $C_0^\infty(\Omega)\times C_0^\infty(\Omega)$. Let us denote by F the completion space of $C_0^\infty(\Omega)\times C_0^\infty(\Omega)$ with the norm in (2.13). It is clear that F is a Hilbert space. Let us multiply the system (2.1) by $q_k(\partial\phi/\partial x_k)$, where $q=(q_1, \cdots, q_k, \cdots, q_n)$ is a C^2 field of vectors satisfying $q=\nu$ in Γ and ν is the exterior normal of Ω. Then we have the following identity:

$$\begin{aligned}\frac{1}{2}\int_\Sigma a(t)q_k\nu_k\left|\frac{\partial\phi}{\partial\nu}\right|^2 d\Sigma &= \left[\left(\phi'(t), q_k\frac{\partial\phi}{\partial x_k}\right)_\Omega\right]_0^T + \frac{1}{2}\int_Q \frac{\partial q_k}{\partial x_k}\{|\phi'|^2 - a(t)|\nabla\phi|^2\}\,dx\,dt \\ &\quad + \int_Q a(t)\frac{\partial q_k}{\partial x_k}\frac{\partial\phi}{\partial x_k}\frac{\partial\phi}{\partial x_j}\,dx\,dt\end{aligned}$$

from which we have

$$\int_Q \left|\frac{\partial\phi}{\partial\nu}\right|^2 d\Sigma \leqq CE(0) \tag{2.14}$$

where C is a generic constant. From this later inequality we can conclude that the norm of F is equivalent to the norm of the space $H_0^1(\Omega)\times L^2(\Omega)$, and since $C_0^\infty(\Omega)$ is dense in both $H_0^1(\Omega)$ and $L^2(\Omega)$, we obtain that $F = H_0^1(\Omega)\times L^2(\Omega)$. Since Λ is a self-adjoint operator from F to F', from Lemma 2.1 and the relations (2.11) and (2.12), we conclude that Λ is an isomorphism. Finally, if we take $\{y_1, -y_0\}$ *in* F', there exists $\{\phi_0, \phi_1\}$ in F such that

$$\Lambda\{\phi_0, \phi_1\} = \{y_1, y_0\}.$$

From (2.12) we have that the function ψ defined by system (2.10) satisfies $\psi(0)=y_0$, $\psi'(0)=y_1$, taking $v=w$ in system (1.1). By uniqueness of solutions for linear hyperbolic systems, we conclude that $y=\psi$, and from (2.10) the result follows. □

REFERENCES

[1] V. KOMORNIK, *Controllabilité exacte en un temps minimal*, C.R. Acad. Sci. Paris, 305 (1987), pp. 605-608.

[2] J. L. LIONS, *Controllabilité exacte des systèmes distribués*, Collège de France, Departement de Mathématiques, #3 Rue d'Ulm, 75005 Paris, 1988.

[3] ———, *Controllabilité exacte des systèmes distribués*, C.R. Acad. Sci. Paris, 302 (1986), pp. 471-475.

[4] ———, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM National Meeting, Boston, MA, July 21-25, 1986.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 503-512, May 1990

A GENERALIZATION OF THE PROXIMAL POINT ALGORITHM*

CU D. HA†

Abstract. The problem considered in this paper is to find a solution to the generalized equation $0 \in T(x, y)$, where T is a maximal monotone operator on the product $H_1 \times H_2$ of two Hilbert spaces H_1 and H_2. We give a generalization of the proximal map and the proximal point algorithm in which the proposed iterative procedure is based on just one variable. Applying to convex programming problems, instead of adding a quadratic term for all variables as in the proximal point algorithm, a quadratic term for a subset of variables is added. This paper proves that under a mild assumption our algorithm has the same convergence properties as the regular proximal point algorithm.

Key words. monotone operator, convex programming, proximal point algorithm, generalized equation

AMS (MOS) subject classifications. 47H05, 49D45, 90C25

1. Introduction. Let H be a real Hilbert space with the inner product $\langle \cdot, \cdot \rangle$ and the induced normed $|\cdot|$. Let $T: H \to H$ be a set-valued map. We define its domain, image, and graph, as follows.

$$\text{Dom}(T) := \{z \in H \mid T(z) \neq \phi\},$$

$$\text{Im}(T) := \bigcup_{z \in H} T(z),$$

and

$$\text{Graph}(T) := \{(z, w) \in H \times H \mid w \in T(z)\}.$$

The inverse T^{-1} of T is the set-valued map defined by

$$z \in T^{-1}(w) \quad \text{if and only if } w \in T(z).$$

The set-valued map T is said to be a monotone operator if

$$\langle z - z', w - w' \rangle \geqq 0 \quad \text{for all } (z, w) \text{ and } (z', w') \text{ in Graph}(T).$$

T is said to be a maximal monotone operator if it is monotone and its graph is not properly contained in the graph of any other monotone operator. The theory and applications of monotone operators have been studied extensively; see, for example, Brezis [3], Dolezal [5], and Aubin and Ekeland [1].

The problem that we are interested in is to find a vector z in H such that

$$0 \in T(z). \tag{1.1}$$

Many problems from mathematical programming, complementarity, mathematical economics and other fields can be formulated as generalized equations (1.1) (Robinson [10]). One specific type of such problems that motivates this paper is the convex programming problem

$$\text{minimize } f(z), \tag{1.2}$$

where f is a closed proper convex function.

* Received by the editors September 16, 1986; accepted for publication (in revised form) June 5, 1989.
† Department of Mathematical Sciences, Virginia Commonwealth University, Richmond, Virginia 23284. Current address, AT&T Bell Laboratories, Holmdel, New Jersey 07733. This work was supported by the Grants-In-Aid Program for Faculty of Virginia Commonwealth University.

Let ∂f be the subgradient of f. It is well known that if f is a closed proper convex function, then ∂f is a maximal monotone operator. Furthermore, $\bar{z}$ solves the minimization problem (1.2) if and only if $\bar{z}$ solves the generalized equation (1.1), where $T = \partial f$.

The proximal point algorithm solves (1.1) when T is maximal monotone. Starting at any point z^0 the algorithm generates a sequence $\{z^k\}$ according to the rule

$$z^{k+1} = P_k(z^k), \tag{1.3}$$

where $P_k(z) := (I + c_k T)^{-1}(z)$ and $\{c_k\}$ is a sequence of positive numbers.

The algorithm is based on the fact that the proximal map P_k is single-valued and nonexpansive (Minty [9]). Rockafellar [13] shows that under certain conditions the sequence $\{z^k\}$ converges to a solution $\bar{z}$ to (1.1). The proximal point algorithm has been investigated further by Luque [8] and Spingarn [14], [15]. It also has been applied to decomposition by Spingarn [16], Ha [6], and Kaneko and Ha [7].

Applying to the convex programming problem (1.2), z^{k+1} satisfies the rule (1.3) if and only if it is the unique optimizer of the problem

$$\text{minimize } f(z) + (1/2c_k)|z - z^k|^2. \tag{1.4}$$

The proximal point algorithm solves (1.2) by iteratively solving (1.4). The advantage of (1.4) over (1.2) is that the function $f(\cdot) + (1/2c_k)|\cdot - z^k|^2$ is strongly convex; consequently, (1.4) has a unique optimal solution. Strong convexity of the objective function is a very important feature if we use methods for solving (1.4) that are based on duality. It is especially significant in the decomposition of large scale problems (see Spingarn [16] and Ha [6]).

Suppose now that z consists of two components $z = (x, y)$, then (x^{k+1}, y^{k+1}) is the optimal solution to the problem

$$\underset{(x,y)}{\text{minimize}}\, f(x, y) + (1/2c_k)|(x, y) - (x^k, y^k)|^2. \tag{1.5}$$

However, in some applications we would like to add a quadratic term of just one variable, say y. That means (x^{k+1}, y^{k+1}) would be an optimal solution to the problem

$$\underset{(x,y)}{\text{minimize}}\, f(x, y) + (1/2c_k)|y - y^k|^2. \tag{1.6}$$

The problem (1.6), in general, does not have as attractive features as (1.5); in particular, the objective function in (1.6) is not as strongly convex in (x, y) as that in (1.5). But in the case that either f is already strongly convex in x for all y or f is separable, i.e., $f(x, y) = f_1(x) + f_2(y)$, $f_1(x)$ is linear and the feasible region is bounded, then (1.6) is easier to solve than (1.5). Moreover, in decomposition methods based on duality, for some problems, (1.6) produces simpler subproblems than (1.5). We illustrate these points by an example.

Example. We consider the following quadratic programming problem:

$$\begin{aligned} &\text{minimize } \langle a_1, x\rangle + (1/2)\langle x, Cx\rangle + \langle a_2, y\rangle \\ &\text{subject to } A_1 x \qquad \leqq b_1 \\ &\qquad\qquad A_2 y \leqq b_2 \\ &\qquad B_1 x + B_2 y \leqq b_3, \end{aligned}$$

where a and b are vectors; A, B, and C are matrices having appropriate dimensions. C is assumed to be positive definite.

Let $f(x, y) = \langle a_1, x\rangle + (1/2)\langle x, Cx\rangle + \langle a_2, y\rangle$. Using the regular proximal point algorithm given by (1.5), the problem to be solved is

$$\begin{aligned} &\text{minimize } f(x, y) + (1/2c_k)|(x, y) - (x^k, y^k)|^2 \\ &\text{subject to } A_1 x \qquad \leqq b_1 \\ &\qquad\qquad A_2 y \leqq b_2 \\ &\qquad B_1 x + B_2 y \leqq b_3. \end{aligned}$$

To decompose the problem we put the coupling constraint into the objective function. The dual problem is

$$\text{maximize } g(u).$$

The dual function $g(u)$ is the optimal objective value of the problem

$$\begin{aligned} &\text{minimize } f(x, y) + (1/2c_k)|(x, y) - (x^k, y^k)|^2 + \langle u, b_3 - B_1 x - B_2 y\rangle \\ &\text{subject to } A_1 x \qquad \leqq b_1 \\ &\qquad\qquad A_2 y \leqq b_2. \end{aligned}$$

The above problem consists of two independent subproblems, one in terms of x variable and one in y variable. We are interested in the subproblem in x variable:

$$\begin{aligned} &\text{minimize } \langle a_1, x\rangle + (1/2)\langle x, Cx\rangle + (1/2c_k)|x - x^k|^2 - \langle u, B_1 x\rangle \\ &\text{subject to } A_1 x \leqq b_1. \end{aligned} \tag{1.7}$$

If we use the generalized proximal point algorithm given by (1.6), then the subproblem in x variable is reduced to

$$\begin{aligned} &\text{minimize } \langle a_1, x\rangle + (1/2)\langle x, Cx\rangle - \langle u, B_1 x\rangle \\ &\text{subject to } A_1 x \leqq b_1. \end{aligned} \tag{1.8}$$

For the problem (1.7) we have to update the matrix of the objective function in each iteration of the proximal point algorithm; meanwhile, the matrix of (1.8) remains constant. That advantage of (1.8) is particularly useful if we use a conjugate direction approach for solving the subproblems. Because the matrix of (1.8) is unchanged, the conjugate directions need not be recomputed for any iteration of the generalized proximal point algorithm.

In this paper we propose an algorithm to solve the generalized equation in two variables

$$0 \in T(x, y).$$

The algorithm is a generalization of the proximal point algorithm. Its iterations are based on just one variable—not both as in the proximal point algorithm. Applying to the convex programming problem (1.2) our generalized proximal point algorithm solves iteratively (1.6) instead of (1.5). In § 2 we define the generalized proximal map as

$$P = (\Pi + cT)^{-1}\Pi$$

where Π is the projection on the second space. We show that under a mild condition, namely, $0 \in \operatorname{int} \operatorname{Im}(T)$, the generalized proximal map P has desirable properties. In §§ 3 and 4 we prove the convergence and the rate of convergence of the generalized

proximal point algorithm. Applications to convex programming problems will be reported in a subsequent paper.

2. Properties of the generalized proximal map. Let $H = H_1 \times H_2$, where H_1 and H_2 are two real Hilbert spaces. For convenience we denote the inner products and the norms in H_1 and H_2 by the same notations $\langle \cdot, \cdot \rangle$ and $|\cdot|$, respectively.

The inner product on H is induced from those on H_1 and H_2, i.e.,

$$\langle (x, y), (u, v) \rangle = \langle x, u \rangle + \langle y, v \rangle$$

for $(x, y) \in H_1 \times H_2$ and $(u, v) \in H_1 \times H_2$.

Let Π be the orthogonal projection of H onto $\{0\} \times H_2$, i.e., $\Pi(x, y) = (0, y)$ for $(x, y) \in H_1 \times H_2$. Let $T: H \to H$ be a maximal monotone operator. We consider the problem of finding $(x, y) \in H_1 \times H_2$ satisfying the generalized equation

$$0 \in T(x, y). \tag{2.1}$$

The generalized proximal point algorithm generates from any point $y^0 \in H_2$ a sequence $\{(x^k, y^k)\}$ by the rule

$$(x^{k+1}, y^{k+1}) \in (\Pi + c_k T)^{-1}(0, y^k), \tag{2.2}$$

where $\{c_k\}$ is a sequence of positive numbers.

We denote

$$P := (\Pi + cT)^{-1}\Pi \tag{2.3}$$

for some positive number c and

$$P_k := (\Pi + c_k T)^{-1}\Pi.$$

Then (2.2) is

$$(x^{k+1}, y^{k+1}) \in P_k(x^k, y^k).$$

Note that if Π is substituted by the identity I then P is the regular proximal map $\bar{P}$ of T. That is $\bar{P} = (I + cT)^{-1}$. The proximal map $\bar{P}$ has several familiar properties:

(i) Dom $\bar{P} = H$.

(ii) $\bar{P}$ is single-valued.

(iii) $\bar{P}$ is nonexpansive, i.e.,

$$|P(z) - P(z')| \leqq |z - z'| \quad \text{for } z, z' \in H.$$

(iv) $\bar{P}(z) = z$ if and only if $0 \in T(z)$.

It is easy to see that P defined by (2.3) does not have any property above. However, we will show that under a simple condition P has property (i) and properties similar to (ii)-(iv).

DEFINITION (Aubin and Ekeland [1, p. 392]). Let A be a monotone operator from X to X, a Hilbert space. It satisfies the L property if for all $w \in \text{Im}(A)$ and $y \in \text{Dom}(A)$ there is a number c such that

$$\inf_{(x, u) \in \text{Graph}(A)} \langle u - w, x - y \rangle \geqq c.$$

The following theorem is proved by Brezis and Haraux [4] but its present form is from Aubin and Ekeland [1, p. 393].

THEOREM. *Let A and B be two monotone operators satisfying*

(i) Dom $(A) \subset$ Dom (B),

(ii) *$A + B$ is maximal monotone, and*

(iii) *B satisfies the L property.*

Then

(i) int Im $(A + B) =$ int $(\text{Im}(A) + \text{Im}(B))$ *and*

(ii) cl Im $(A+B) = \text{cl}(\text{Im}(A)+\text{Im}(B))$, *where* int *and* cl *denote the topological interior and closure, respectively.*

PROPOSITION 1. *Suppose* $0 \in \text{int Im}(T)$. *Then*

$$\{0\} \times H_2 \subset \text{int Im}(\Pi + cT) \tag{2.4}$$

and Dom $(P) = H$.

Proof. To use the theorem of Brezis and Haraux with $A = cT$ and $B = \Pi$ we need to verify its three conditions.

Since Dom $(\Pi) = H$ the first condition is trivial. We also have

$$\text{Dom}(cT) \cap \text{int}(\text{Dom}\,\Pi) = \text{Dom}(cT) \neq \phi.$$

Therefore, $cT + \Pi$ is maximal monotone (Rockafellar [12]).

We now show that Π satisfies the L property. Let $x = (x_1, x_2)$, $y = (y_1, y_2)$ and $z = (z_1, z_2)$ be three vectors in $H_1 \times H_2$.

Let

$$u = \Pi(x_1, x_2) = (0, x_2),$$

$$v = \Pi(y_1, y_2) = (0, y_2),$$

and

$$w = \Pi(z_1, z_2) = (0, z_2).$$

We have

$$\begin{aligned}\langle u, x-y\rangle + \langle v, y-z\rangle + \langle w, z-x\rangle &= \langle x_2, x_2-y_2\rangle + \langle y_2, y_2-z_2\rangle + \langle z_2, z_2-x_2\rangle \\ &= \frac{1}{2}(|x_2-y_2|^2 + |y_2-z_2|^2 + |z_2-x_2|^2) \geqq 0.\end{aligned}$$

Define $c := \langle w-v, y-z\rangle$. Then $\langle u-w, x-y\rangle \geqq c$ for all $(x, u) \in \text{graph}(\Pi)$. By the theorem of Brezis and Haraux

$$\begin{aligned}\text{int}(\text{Im}(cT+\Pi)) &= \text{int}(\text{Im}(cT) + \text{Im}(\Pi)) \\ &= \text{int}(\text{Im}(cT) + \{0\} \times H_2).\end{aligned}$$

Some $0 \in \text{int Im}(T) = \text{int Im}(cT)$ we have

$$\{0\} \times H_2 \subset \text{int}(\text{Im}(cT+\Pi)).$$

Let $(x, y) \in H$, then by (2.4) there is $(u, v) \in H$ such that

$$\Pi(x, y) = (0, y) \in (\Pi + cT)(u, v).$$

That implies

$$(u, v) \in (\Pi + cT)^{-1}\Pi(x, y) = P(x, y).$$

Therefore

$$\text{Dom } P = H.$$

PROPOSITION 2.

(i) *If* $(u_i, v_i) \in P(x, y)$ *for* $i = 1, 2$ *then* $v_1 = v_2$.

(ii) *If* $(u_i, v_i) \in P(x_i, y_i)$ *for* $i = 1, 2$ *then* $|v_1 - v_2| \leqq |y_1 - y_2|$.

(iii) $0 \in T(x, y)$ *if and only if* $(x, y) \in P(x, y)$.

Proof.

(i) $(u_i, v_i) \in P(x, y)$
$\Rightarrow (0, y) \in (\Pi + cT)(u_i, v_i)$
$\Rightarrow (0, y - v_i) \in cT(u_i, v_i)$ for $i = 1, 2$.

By monotonicity of cT we have

$$\langle (0, y - v_1) - (0, y - v_2), (u_1, v_1) - (u_2, v_2) \rangle \geqq 0$$
$$\Rightarrow \langle v_2 - v_1, v_1 - v_2 \rangle \geqq 0.$$

Therefore $v_1 = v_2$.

(ii) $(u_i, v_i) \in P(x_i, y_i)$
$\Rightarrow (0, y_i - v_i) \in cT(u_i, v_i)$ for $i = 1, 2$
$\Rightarrow \langle (y_1 - v_1) - (y_2 - v_2), v_1 - v_2 \rangle \geqq 0.$

We have

$$|y_1 - y_2|^2 = |y_1 - y_2 - v_1 + v_2 + v_1 - v_2|^2$$
$$= |y_1 - y_2 - v_1 + v_2|^2 + |v_1 - v_2|^2 + 2\langle y_1 - y_2 - v_1 + v_2, v_1 - v_2 \rangle.$$

Hence

$$|y_1 - y_2|^2 \geqq |y_1 - y_2 - v_1 + v_2|^2 + |v_1 - v_2|^2 \tag{2.5}$$

and

$$|y_1 - y_2| \geqq |v_1 - v_2|.$$

(iii) $(x, y) \in P(x, y)$
$\Leftrightarrow (0, y) \in (\Pi + cT)(x, y)$
$\Leftrightarrow (0, y) - (0, y) \in cT(x, y)$
$\Leftrightarrow 0 \in T(x, y)$.

Although P does not have as good properties as the regular proximal map, from Proposition 2 we observe that P is well behaved enough for our purpose. While P is neither single-valued nor nonexpansive, the second component is uniquely determined and nonexpansive. Moreover, the generalized equation (2.1) can be converted to a fixed point problem.

3. Convergence of the generalized proximal point algorithm. For practical purposes (x^{k+1}, y^{k+1}) should be obtained according to some approximation criteria rather than exactly as in (2.2). Following Rockafellar [13] we consider two approximation criteria

$$|(x^{k+1}, y^{k+1}) - (u^{k+1}, v^{k+1})| \leqq \varepsilon_k, \qquad \sum_{k=0}^{\infty} \varepsilon_k < \infty \tag{A}$$

and

$$|(x^{k+1}, y^{k+1}) - (u^{k+1}, v^{k+1})| \leqq \delta_k |(x^{k+1}, y^{k+1}) - (x^k, y^k)| \tag{B}$$

and

$$|(y^{k+1} - v^{k+1})| \leqq \delta_k |y^{k+1} - y^k|, \qquad \sum_{k=0}^{\infty} \delta_k < \infty,$$

where

$$(u^{k+1}, v^{k+1}) \in P(x^k, y^k). \tag{3.1}$$

Other approximation criteria, as in Auslender [2], will be discussed in a subsequent paper that applies the generalized proximal point algorithm to convex programming problems.

THEOREM 1. *Let* $\{c_k\}$ *be a sequence of positive numbers bounded away from zero. Suppose* $0 \in \operatorname{int} \operatorname{Im}(T)$. *Let* $\{(x^k, y^k)\}$ *be a sequence generated under criterion* (A). *Then* $\{(x^k, y^k)\}$ *is bounded and any of its weak cluster point is a solution to* (2.1). *Moreover,* $\{y^k\}$ *converges weakly to* $\bar{y}$, *a second component of a solution to* (2.1).

Proof. Let $(\bar{x}, \bar{y})$ be a solution to (2.1), then

$$(\bar{x}, \bar{y}) \in P(\bar{x}, \bar{y}).$$

Similarly to (2.5) the relation above and the definition (3.1) of (u^{k+1}, v^{k+1}) yield

$$|y^k - \bar{y}|^2 \geqq |y^k - v^{k+1}|^2 + |v^{k+1} - \bar{y}|^2. \tag{3.2}$$

Hence

$$|v^{k+1} - \bar{y}| \leqq |y^k - \bar{y}|.$$

From

$$|y^{k+1} - \bar{y}| \leqq |y^{k+1} - v^{k+1}| + |v^{k+1} - \bar{y}|$$

it implies that

$$|y^{k+1} - \bar{y}| \leqq \varepsilon_k + |y^k - \bar{y}|.$$

Because $\sum_{k=0}^{\infty} \varepsilon_k < \infty$ the sequence $\{y^k\}$ is bounded and $\lim_{k\to\infty} |y^k - \bar{y}|$ exists and is finite.

We now show that $|y^k - v^{k+1}| \to 0$ as $k \to \infty$. By (3.2) we have

$$\begin{aligned}
|y^k - v^{k+1}|^2 - |y^k - \bar{y}|^2 + |y^{k+1} - \bar{y}|^2 &\leqq |y^{k+1} - \bar{y}|^2 - |v^{k+1} - \bar{y}|^2 \\
&= \langle y^{k+1} - v^{k+1}, (y^{k+1} - \bar{y}) + (v^{k+1} - \bar{y}) \rangle \\
&\leqq |y^{k+1} - v^{k+1}|(|y^{k+1} - \bar{y}| + |v^{k+1} - \bar{y}|) \\
&\leqq \varepsilon_k(|y^{k+1} - \bar{y}| + |y^k - \bar{y}|) \\
&\leqq 2\varepsilon_k(s + |\bar{y}|),
\end{aligned}$$

where s is a positive number such that

$$|y^k| \leqq s \quad \text{for all } k.$$

Consequently

$$|y^k - v^{k+1}|^2 \leqq |y^k - \bar{y}|^2 - |y^{k+1} - \bar{y}|^2 + 2\varepsilon_k(s + |\bar{y}|).$$

Hence

$$|y^k - v^{k+1}| \to 0 \quad \text{as } k \to \infty.$$

Since $\{c_k\}$ is bounded away from zero, we also have

$$(1/c_k)(y^k - v^{k+1}) \to 0 \quad \text{as } k \to \infty. \tag{3.3}$$

The assumption $0 \in \operatorname{int} \operatorname{Im}(T)$ implies that there exist positive numbers α and ε such that

$$|z| \geqq \alpha \Rightarrow |w| \geqq \varepsilon \text{ whenever } w \in T(z)$$

(Rockafellar [13]) or equivalently for any $w \in B(0, \varepsilon)$, $T^{-1}(w) \subset B(0, \alpha)$, where $B(0, \varepsilon)$ and $B(0, \alpha)$ denote the ball of center 0 and radius ε and α, respectively. In other words T^{-1} is locally bounded at 0. By (3.3) there is an integer K such that

$$(0, (1/c_k)(y^k - v^{k+1})) \in B(0, \varepsilon) \text{ for } k \geqq K.$$

That implies

$$(u^{k+1}, v^{k+1}) \in T^{-1}(0, (1/c_k)(y^k - v^{k+1})) \subset B(0, \alpha) \quad \text{for } k \geqq K.$$

Therefore $\{(u^k, v^k)\}$ is bounded and so is $\{(x^k, y^k)\}$.

Let $\{(x^{k_j}, y^{k_j})\}$ be a subsequence that converges weakly to (x^*, y^*). We observe that $\{(u^{k_j}, v^{k_j})\}$ also converges weakly to (x^*, y^*) and

$$(u^{k_j}, v^{k_j}) \in T^{-1}(0, (1/c_{k_j-1})(y^{k_j-1} - v^{k_j})).$$

We know that the graph of a maximal monotone operator is weakly-strongly closed (Aubin and Ekeland [1, p. 379]). Therefore at the limit

$$(x^*, y^*) \in T^{-1}(0, 0)$$

or (x^*, y^*) is a solution to (2.1).

Let $(\tilde{x}, \tilde{y})$ be another weak cluster point of $\{(x^k, y^k)\}$; we will show that $\tilde{y} = y^*$. This proof is similar to the one given by Rockafellar [13, p. 885]. Using the same argument as above, $(\tilde{x}, \tilde{y})$ is a solution to (2.1). Likewise

$$\mu_1 = \lim_{k\to\infty} |y^k - y^*|$$

and

$$\mu_2 = \lim_{k\to\infty} |y^k - \tilde{y}|$$

exist and are finite. From

$$|y^k - \tilde{y}|^2 = |y^k - y^*|^2 + |y^* - \tilde{y}|^2 + 2\langle y^k - y^*, y^* - \tilde{y}\rangle$$

we see that the limit of $\langle y^k - y^*, y^* - \tilde{y}\rangle$ as $k \to \infty$ must exist. That limit has to be zero because y^* is a weak cluster point of $\{y^k\}$. Hence, at the limit

$$\mu_2 = \mu_1 + |y^* - \tilde{y}|^2.$$

Reversing the role of $\tilde{y}$ and y^* we also have

$$\mu_1 = \mu_2 + |y^* - \tilde{y}|^2.$$

That implies $\tilde{y} = y^*$.

4. Rate of convergence. This section follows closely § 3 in Rockafellar [13]. In other words we obtain the same linear rate and if $c_k \uparrow \infty$ the rate is superlinear. Also under certain conditions we have finite convergence.

DEFINITION (Rockafellar [13]). A set-valued map A is said to be Lipschitz continuous at w_0 with modulus $a \geqq 0$ if $A(w_0)$ is single-valued, i.e., $A(w_0) = \{z_0\}$ and for some $\tau > 0$ we have

$$|z - z_0| \leqq a|w - w_0| \quad \text{whenever } z \in A(w) \text{ and } |w - w_0| < \tau.$$

THEOREM 2. *Suppose T^{-1} is Lipschitz continuous at 0 with modulus a. Let $\{(x^k, y^k)\}$ be any sequence generated under Criterion (B) with $\{c_k\}$ nondecreasing ($c_k \uparrow c \leqq \infty$). Assume that $\{(x^k, y^k)\}$ is bounded.*

Let

$$\mu_k := a/(a^2 + c_k^2)^{1/2}.$$

Then $\{(x^k, y^k)\}$ converges strongly to $(\bar{x}, \bar{y})$, the unique solution to (2.1). Moreover, there is an integer K such that

$$|y^{k+1} - \bar{y}| \leqq \theta_k |y^k - \bar{y}| \quad \textit{for all } k \geqq K$$

where $\theta_k := (\mu_k + \delta_k)/(1 - \delta_k)$ for all $k \geqq K$.

Proof. By the definition of Lipschitz continuity, T^{-1} is locally bounded at 0, so $0 \in \operatorname{int} \operatorname{Im}(T)$ (Rockafellar [11]). The boundedness of $\{(x^k, y^k)\}$ allows us to use Theorem 1 with

$$\varepsilon_k = \delta_k |(x^{k+1}, y^{k+1}) - (x^k, y^k)|.$$

By the definition of (u^{k+1}, v^{k+1})

$$(u^{k+1}, v^{k+1}) \in T^{-1}(0, (1/c_k)(y^k - v^{k+1})).$$

By the proof of Theorem 1

$$(1/c_k)(y^k - v^{k+1}) \to 0 \quad \text{as } k \to \infty.$$

Consequently, there is an integer K such that

$$|(1/c_k)(y^k - v^{k+1})| < \tau \quad \text{for all } k \geqq K.$$

That implies

$$|(u^{k+1}, v^{k+1}) - (\bar{x}, \bar{y})| \leqq a|(0, (1/c_k)(y^k - v^{k+1}))|. \tag{4.1}$$

We have

$$\begin{aligned} |(x^{k+1}, y^{k+1}) - (\bar{x}, \bar{y})| &\leqq |(x^{k+1}, y^{k+1}) - (u^{k+1}, v^{k+1})| + |(u^{k+1}, v^{k+1}) - (\bar{x}, \bar{y})| \\ &\leqq 2\delta_k s + a|(0, (1/c_k)(y^k - v^{k+1}))|, \end{aligned}$$

where s is a number such that

$$|(x^k, y^k)| \leqq s \quad \text{for all } k.$$

Hence $\{(x^k, y^k)\}$ converges strongly to $(\bar{x}, \bar{y})$.

We have from (4.1)

$$|v^{k+1} - \bar{y}| \leqq (a/c_k)|y^k - v^{k+1}|$$

and from (3.2)

$$|v^{k+1} - \bar{y}|^2 + |y^k - v^{k+1}|^2 \leqq |y^k - \bar{y}|^2.$$

These imply

$$(1 + (a/c_k)^2)|v^{k+1} - \bar{y}|^2 \leqq (a/c_k)^2 |y^k - \bar{y}|^2$$

or

$$|v^{k+1} - \bar{y}| \leqq (a/(a^2 + c_k^2)^{1/2})|y^k - \bar{y}| = \mu_k |y^k - \bar{y}|.$$

On the other hand,

$$|y^{k+1} - \bar{y}| \leqq |y^{k+1} - v^{k+1}| + |v^{k+1} - \bar{y}|$$

and

$$|v^{k+1} - y^{k+1}| \leqq \delta_k |y^{k+1} - y^k| \leqq \delta_k |y^{k+1} - \bar{y}| + \delta_k |\bar{y} - y^k|.$$

Hence

$$|y^{k+1} - \bar{y}| \leqq \delta_k |y^{k+1} - \bar{y}| + \delta_k |y^k - \bar{y}| + \mu_k |y^k - \bar{y}|$$

or

$$(1 - \delta_k)|y^{k+1} - \bar{y}| \leqq (\delta_k + \mu_k)|y^k - \bar{y}|.$$

That gives

$$|y^{k+1}-\bar{y}| \leqq \theta_k |y^k-\bar{y}| \quad \text{where } \theta_k = (\delta_k+\mu_k)/(1-\delta_k).$$

Theorem 2 shows that the rate of convergence is linear and if $c_k \to \infty$ as $k \to \infty$ then $\mu_k \to 0$ and $\theta_k \to 0$, so the rate is superlinear. In the next theorem we show that for a special case the convergence is finite.

THEOREM 3. *Let $\{(x^k, y^k)\}$ be any sequence generated by the generalized proximal point algorithm under criterion (A) or (B) with $\{c_k\}$ bounded away from zero. We assume that $\{(x^k, y^k)\}$ is bounded if criterion (B) is used. Suppose that there is $(\bar{x}, \bar{y})$ such that $0 \in \operatorname{int} T(\bar{x}, \bar{y})$. Then*

$$(\bar{x}, \bar{y}) = (u^{k+1}, v^{k+1}) \text{ for } k \text{ sufficiently large.}$$

In particular, the generalized proximal point algorithm in its exact form $((x^{k+1}, y^{k+1}) \in P(x^k, y^k))$ gives convergence to $(\bar{x}, \bar{y})$ in a finite number of iterations from any starting point.

Proof. Rockafellar [13] shows that if $0 \in \operatorname{int} T(\bar{x}, \bar{y})$, then T^{-1} is single-valued and constant on a neighborhood of 0, i.e., there exists $\varepsilon > 0$ such that if $|w| < \varepsilon$ then $T^{-1}(w) = (\bar{x}, \bar{y})$. By Theorem 1, $(1/c_k)(y^k - v^{k+1}) \to 0$ as $k \to \infty$ and $(u^{k+1}, v^{k+1}) \in T^{-1}(0, (1/c_k)(y^k - v^{k+1}))$. Therefore for k sufficiently large

$$(0, (1/c_k)(y^k - v^{k+1})) \in B(0, \varepsilon).$$

Consequently $(u^{k+1}, v^{k+1}) = (\bar{x}, \bar{y})$ for k sufficiently large.

Applications of the generalized proximal point algorithm to convex programming problems will be discussed in a subsequent paper.

REFERENCES

[1] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley Interscience, New York, 1984.

[2] A. AUSLENDER, *Numerical methods for nondifferentiable convex optimization*, Math. Programming Stud., 30 (1987), pp. 102-126.

[3] H. BREZIS, *Operateurs Maximaux Monotones*, North Holland, Amsterdam, 1973.

[4] H. BREZIS AND A. HARAUX, *Image d'une somme d'operateurs monotones et applications*, Israel Journal of Mathematics, 23 (1976), pp. 165-186.

[5] V. DOLEZAL, *Monotone Operators and Applications in Control and Network Theory*, Elsevier, Amsterdam, 1979.

[6] C. D. HA, *Algorithms to solve large scale structured convex programming problems*, Ph.D. dissertation, Department of Industrial Engineering, University of Wisconsin, Madison, 1980.

[7] I. KANEKO AND C. D. HA, *A decomposition procedure for large-scale optimal plastic design problems*, Internat. J. Numer. Meth. Engrg., 19 (1983), pp. 873-889.

[8] F. J. LUQUE, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM J. Control Optim., 22 (1984), pp. 277-293.

[9] G. J. MINTY, *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J., 29 (1962), pp. 341-346.

[10] S. M. ROBINSON, *Generalized equations and their solutions, part I: Basic theory*, Math. Programming Stud., 10 (1979), pp. 128-141.

[11] R. T. ROCKAFELLAR, *Local boundedness of nonlinear, monotone operators*, Michigan Math. J., 16 (1969), pp. 397-407.

[12] ———, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75-88.

[13] ———, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877-898.

[14] J. E. SPINGARN, *Submonotone mappings and the proximal point algorithm*, Numer. Funct. Anal. Optim., 4 (1982), pp. 123-150.

[15] ———, *Partial inverse of a monotone operator*, Appl. Math. Optim., 10 (1983), pp. 247-265.

[16] ———, *Applications of the method of partial inverses to convex programming: Decomposition*, Math. Programming, 32 (1985), pp. 199-223.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 513-527, May 1990

CONTINUOUS-TIME STOCHASTIC ADAPTIVE TRACKING— ROBUSTNESS AND ASYMPTOTIC PROPERTIES*

HAN-FU CHEN† AND LEI GUO†

Abstract. Adaptive estimation and control problems are considered for continuous-time stochastic systems containing both modeled and unmodeled dynamics. The least squares method is used to estimate unknown parameters included in the modeled part, which are used to update an adaptive control law. It is shown that both the estimation error and the tracking error are bounded, and that the bounds are proportional to constants dominating the unmodeled dynamics. Moreover, convergence rates of the tracking errrors are established in the case where no unmodeled dynamics exist.

Key words. continuous-time stochastic system, adaptive tracking, least squares, robustness, unmodeled dynamics

AMS(MOS) subject classifications. 93C40, 93E12

1. Introduction. In recent years, much attention has been devoted to the analysis of adaptive algorithms when unmodeled dynamics are contained in the system. It is known that (see, e.g., [1]-[3]) unmodeled dynamics or even small disturbances may cause instability in many adaptive algorithms when precautions are not taken. This inspired the study of robust adaptive control where the primary purpose is to maintain stability of the closed-loop system under violations of ideal assumptions. There is already a vast literature on this topic, especially in the deterministic framework (e.g., [4]-[6]).

In the stochastic case, robustness results are much more difficult to obtain. This results from the following "stochastic features": (i) a priori upper bounds for the noise sequence are usually not available, (ii) optimal or at least close to optimal rejection of the noise effects is required, and (iii) traditionally used supermartingale methods fail due to unmodeled dynamics. An initial attempt toward robustness analysis for discrete-time stochastic adaptive systems was made in [7], where an a priori assumption on the input-output data was required. This assumption was later removed in [8] for a large class of stochastic systems represented by a full ARMAX model plus unmodeled dynamics.

While discrete-time adaptive theory is well developed, the corresponding continuous-time analogue becomes a natural concern. There is no doubt that results of this kind are interesting and important in many situations. Unfortunately, it seems that they have received less attention in the literature, and that only some initial works in the adaptive estimation aspect are available (see, e.g., [9]-[12]).

In this paper, we consider both estimation and control problems for stochastic systems described by stochastic differential/integral equations. The adaptive control law is defined based on a continuous-time analogue of the least-squares estimation algorithm. We show the following:

(i) That the least squares method has some degree of robustness when unmodeled dynamics are contained in the model, provided that the system is "persistently excited."

(ii) That the closed-loop adaptive system is stable, with a tracking error upper bound. This bound implies that the tracking error will decrease when upper bounds on the unmodeled dynamics decrease.

* Received by the editors January 21, 1987; accepted for publication (in revised form) August 11, 1989.

† Institute of Systems Science, Academia Sinica, Beijing, People's Republic of China. This project was supported by the National Natural Science Foundation of China and the TWAS research grant 87-43.

(iii) That if there are no unmodeled dynamics, then the least squares estimation results parallel those obtained in the discrete-time case (see, e.g., [13], [14]); furthermore, in the present paper we provide a precise convergence rate for the tracking error, which in the discrete-time case still remains a standing issue.

We state here that the above-mentioned results are established under the assumption that the strong solution of the stochastic differential equations describing the closed-loop system exists. And for the time being, we know of no way to verify or sidestep this assumption. However, we believe that many of the ideas, techniques, and results presented in this paper are necessary preliminaries for future study.

2. The system description. Let $\{F_t\}$ be a family of nondecreasing σ-algebras defined on a probability space (Ω, F, P), and let the system to be considered be described by the following stochastic differential/integral equation:

$$\begin{aligned}(1)\qquad [I+\mu_1 SH_1(S)]A(S)y_t &= [I+\mu_2 H_2(S)]SB(S)u_t + [I+\mu_3 SH_3(S)]C(S)v_t \\ &\quad + \mu_4 S\xi_t(y,u), \qquad t\geqq 0, \quad y_0=0, \quad u_0=0, \quad \xi_0=0\end{aligned}$$

where s denotes the integral operator (e.g., $Sy_t=\int_0^t y_z\,dz$), and y_t and u_t adapted to $\{F_t\}$ are m-dimensional output and l-dimensional input, respectively. The quantities μ_i, $i=1,\cdots,4$, are small constants, $H_i(S)$, $i=1,2,3$, are unmodeled matrix transfer functions, and $\xi_t(y,u)$, dependent on the previous observation $\{y_s, u_2, 0\leqq s\leqq t\}$, is an unknown nonanticipative measurable process characterizing the unmodeled dynamics. Finally, v_t is the system noise that is generated via a known filter $D^{-1}(S)$ from a standard Wiener process (w_t, F_t):

$$(2)\qquad D(S)v_t = w_t, \qquad t\geqq 0.$$

Assume that $A(S)$, $B(S)$, and $C(S)$ are matrix polynomials in S, with unknown coefficients but known upper bounds for the true orders:

$$(3)\qquad A(S)=I+A_1S+\cdots+A_pS^p, \qquad p\geqq 0,$$

$$(4)\qquad B(S)=B_1+B_2S+\cdots+B_qS^{q-1}, \qquad q\geqq 1,$$

$$(5)\qquad C(S)=I+C_1S+\cdots+C_rS^r, \qquad r\geqq 1,$$

$$(6)\qquad D(S)=I+D_1S+\cdots+D_rS^r.$$

Note that (1) may be rewritten in the form

$$(7)\qquad A(S)y_t = SB(S)u_t + C(S)v_t + \eta_t,$$

$$\begin{aligned}(8)\qquad \eta_t &= \mu_4 S\xi_t(y,u) - \mu_1 SH_1(S)A(S)y_t + \mu_2 SH_2(S)B(S)u_t \\ &\quad + \mu_3 SH_3(S)C(S)v_t.\end{aligned}$$

We remark that, if the unmodeled dynamics are removed, i.e., $\eta_t=0$, for all $t\geqq 0$, then the model (7) is reduced to the one considered in [9]-[12]. Clearly, in this case model (7) may be rewritten in the standard linear state space form, and the output process $\{y_t\}$ can be uniquely determined by the process $\{u_t, w_t\}$. In the general case, it is natural to assume that $\{y_t\}$ can also be determined by $\{u_t, w_t\}$ via (7)-(8).

We denote the collection of unknown matrix coefficients of $A(S)$, $B(S)$, and $C(S)$ by θ:

$$(9)\qquad \theta^\tau = [-A_1\cdots -A_p \quad B_1\cdots B_q \quad C_1\cdots C_r].$$

In the sequel, θ is estimated by the continuous-time extended least square algorithm [10]-[12]:

$$d\theta_t = P_t\phi_t D(S)(dy_t^\tau - \phi_t^\tau \theta_t\, dt), \qquad \theta_0 = 0, \tag{10}$$

$$dP_t = -P_t\phi_t\phi_t^\tau P_t\, dt, \quad P_0 = aI \quad (a = \dim \text{ of } \phi_t), \tag{11}$$

$$\phi_t = [y_t^\tau, Sy_t^\tau \cdots S^{p-1}y_t^\tau, u_t^\tau, Su_t^\tau \cdots S^{q-1}u_t^\tau, \hat{v}_t^\tau \cdots S^{r-1}\hat{v}_t^\tau]^\tau, \tag{12}$$

$$\hat{v}_t = y_t - S\theta_t^\tau \phi_t. \tag{13}$$

Obviously, if $r = 0$, then (10) and (11) can be expressed as

$$\theta_t = P_t \int_0^t \phi_s\, dy_s^\tau + P_t(P_0)^{-1}\theta_0, \tag{14}$$

$$P_t = \left(\int_0^t \phi_s\phi_s^\tau\, ds + a^{-1}\right)^{-1}, \tag{15}$$

and the right-hand side of (14) is completely determined by the observations $\{u_s, y_s, s \leqq t\}$.

In the general $r > 0$ case, however, the regressor ϕ_t depends on $\{\theta_s, s \leqq t\}$. Then (10) and (11) constitute a system of nonlinear stochastic differential equations for θ_t. The existence of the solution is far from obvious since the typical Lipschitz condition, which plays a vital role in the standard theory of stochastic differential equations (see, [15, Chap. 4], for example), is hard to verify in the present case. For that study, the introduction of new techniques seems to be necessary, although our differential equations are well motivated.

Henceforth, we assume that the stochastic differential/integral equation (10)-(11) has a unique strong solution $\{\theta_t, t \geqq 0\}$ in the sense of [15, pp. 127].

Set

$$\phi_t^0 = [y_t^\tau, Sy_t^\tau \cdots S^{p-1}y_t^\tau, u_t^\tau, Su_t^\tau \cdots S^{q-1}u_t^\tau, v_t^\tau \cdots S^{r-1}v_t^\tau]^\tau, \tag{16}$$

$$\tilde{\phi}_t = [0 \cdots 0, 0 \cdots 0, \tilde{v}_t^\tau \cdots S^{r-1}\tilde{v}_t^\tau]^\tau \qquad \tilde{v}_t = v_t - \hat{v}_t, \tag{17}$$

$$Y_t = [y_t^\tau \cdots S^{p-1}y_t^\tau]^\tau, \qquad U_t = [u_t^\tau \cdots S^{q-1}u_t^\tau]^\tau, \tag{18}$$

$$V_t = [v_t^\tau \cdots S^{r-1}v_t^\tau]^\tau, \quad \hat{V}_t = [\hat{v}_t^\tau \cdots S^{r-1}\hat{v}_t^\tau]^\tau, \quad \tilde{V}_t = V_t - \hat{V}_t. \tag{19}$$

Then it follows that

$$\phi_t = [Y_t^\tau, U_t^\tau, \hat{V}_t^\tau]^\tau, \quad \phi_t^0 = [Y_t^\tau, U_t^\tau, V_t^\tau]^\tau, \quad \tilde{\phi}_t = [0, 0, \tilde{V}_t^\tau]^\tau. \tag{20}$$

Furthermore, we set

$$F_d = \begin{bmatrix} -D_1 & \cdots & \cdots & -D_r \\ I & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & I & 0 \end{bmatrix}, \qquad F_c = \begin{bmatrix} -C_1 & \cdots & \cdots & -C_r \\ I & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & I & 0 \end{bmatrix}. \tag{21}$$

By use of these kinds of matrices it is easy to represent an input-output equation in state space form. For example, from (2) and (19) we may write V_t as

$$dV_t = F_d V_t\, dt + [I, 0 \cdots 0]^\tau\, dw_t. \tag{22}$$

In the sequel, similar representations will be used without additional explanation.

On the unmodeled dynamics η_t, we make an assumption similar to that used for discrete-time systems in [8].

Assumption 1. There is a real number $\varepsilon \geqq 0$ such that

$$\int_0^t \|\dot{\eta}_s\| \, ds \leqq \varepsilon r_t, \qquad t \geqq 0 \tag{23}$$

where

$$\dot{\eta}_t = \mu_4 \xi_t(y, u) - \mu_1 H_1(S)A(S)y_t + \mu_2 H_2(S)B(S)u_t + \mu_3 H_3(S)C(S)v_t \tag{24}$$

and

$$r_t = e + \int_0^t \|\phi_s\|^2 \, ds. \tag{25}$$

We also need the following condition on the noise model, which in the discrete-time case is a standard assumption.

Assumption 2. $D(S)$ is stable and the transfer matrix $D(S)C^{-1}(S) - I/2$ is strictly positive real.

At first sight, Assumption 1 is somewhat hard to understand and rather restrictive. However, the following examples show that there is at least one substantial and important class of dynamical systems that does satisfy this condition.

Example 1. Let the single-input and single-output system be described by the following system with additive noise:

$$y_t = G_0(S)[I + \mu G_1(S)]Su_t + v_t, \tag{26}$$

where $G_0(S) = B(S)/A(S)$ represents the nominal transfer function, whereas $G_1(S)$ is the unmodeled transfer function and is assumed to be stable and proper.

When the additive noise v_t is identically equal to zero, then the system is reduced to the deterministic one, and it coincides with the model considered (e.g., [6]) in the robustness analysis for deterministic systems.

Putting the expression for $G_0(S)$ into (26) leads to

$$A(S)y_t = B(S)u_t + \mu S G_1(S)B(S)u_t + A(S)v_t.$$

Comparing this to (7) shows that in the present case

$$\dot{\eta}_t = \mu G_1(S)B(S)u_t. \tag{27}$$

We now prove that Assumption 1 is satisfied for the system (26). For this the following auxiliary result is needed. We formulate it as a lemma, as it will also be used in the proof of the main results to follow.

LEMMA 1. *Let $E(S)$ and $F(S)$ be matrix polynomials in the integral operator S, such that the transfer matrix $F(S)E^{-1}(s)$ is stable and proper. Then*

$$\int_0^t \|F(S)E^{-1}(S)x_z\|^2 \, dz \leqq c \int_0^t \|x_z\|^2 \, dz$$

for any square integrable function $\{x_t\}$, where c is a constant depending on $E(s)$ and $F(S)$ only.

Proof. Let us write

$$E(S) = I + ES + \cdots + E_d S^d, \qquad F(S) = I + F_1 S + \cdots + F_d S^d$$

and set

$$z_t = E^{-1}(S)x_t, \qquad Z_t = [z_t^\tau \cdots S^{d-1}z_t^\tau]^\tau.$$

Similar to (22) we have

$$Z_t = F_e S Z_t + [x_t^\tau, 0 \cdots 0]^\tau \tag{28}$$

with

$$F_e = \begin{bmatrix} -E_1 & \cdots & \cdots & -E_d \\ I & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & I & 0 \end{bmatrix}.$$

The above linear differential equation has the solution

$$Z_t = F_e \int_0^t \exp\{F_e(t-s)\}[x_s^\tau, 0 \cdots 0]^\tau \, ds + [x_t^\tau, 0 \cdots 0]^\tau.$$

Since F_e is stable, there are constants $c_1 \geqq 1$ and $\rho > 0$ such that

$$\|\exp\{F_e t\}\| \leqq c_1 e^{-\rho t} \quad \forall t \geqq 0$$

where here and hereafter c_i, $i = 1, 2, \cdots$, denote constants.

It then follows that

$$\begin{aligned} \int_0^t \|Z_z\|^2 \, dz &\leqq 2\|F_e\|^2 \int_0^t \left\{ \left\| \int_0^z \exp[F_e(z-s)][x_s^\tau, 0 \cdots 0]^\tau \, ds \right\|^2 + \|x_z\|^2 \right\} dz \\ &\leqq 2(c_1)^2 \|F_e\|^2 \\ &\quad \cdot \int_0^t \left\{ \int_0^z \exp[-\rho(z-s)] \, ds \int_0^z \exp[-\rho(z-s)] \|x_s\|^2 \, ds + \|x_z\|^2 \right\} dz \\ &\leqq 2(c_1)^2 \rho^{-1} \|F_e\|^2 \int_0^t \left\{ \int_0^z \exp[-\rho(z-s)] \|x_s\|^2 \, ds + \|x_z\|^2 \right\} dz \\ &\leqq 2(c_1)^2 \rho^{-1} \|F_e\|^2 \left\{ \int_0^t \|x_z\|^2 \, dz + \rho^{-1} \int_0^t \|x_s\|^2 \, ds \right\} \\ &\leqq 2(c_1)^2 \rho^{-1} \|F_e\|^2 (1 + \rho^{-1}) \int_0^t \|x_s\|^2 \, ds. \end{aligned} \tag{29}$$

Furthermore, by (28) it follows that

$$SZ_t = (F_e)^{-1} \left\{ Z_t - \begin{bmatrix} x_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\}$$

and

$$\int_0^t \|SZ_z\|^2 \, dz \leqq c_2 \int_0^t \|x_z\|^2 \, dz \tag{30}$$

by (29).

Finally, the lemma follows from (29) and (30):

$$\int_0^t \|F(S)E^{-1}(S)x_s\|^2\,ds = \int_0^t \|z_s + F_1 S z_s + \cdots + F_d S^d z_s\|^2\,ds$$

$$= \int_0^t \|[I, 0 \cdots 0]z_s + [F_1 \cdots F_d]SZ_s\|^2\,ds$$

$$\leqq c\int_0^t \|x_s\|^2\,ds. \qquad \square$$

We now turn back to show that $\dot{\eta}_t$ given by (27) satisfies Assumption 1.
Set

$$x_t = \mu B(S)u_t.$$

We then have

$$\int_0^t \|x_s\|^2\,ds \leqq \mu^2 \int_0^t \|B_1 u_s + \cdots + B_q S^{q-1}u_s\|^2\,ds \leqq \mu^2 c_3 r_t,$$

where r_t is defined by (25) and c_3 is a constant.

Finally, applying Lemma 1 to (27), we find that

$$\int_0^t \|\dot{\eta}_s\|^2\,ds \leqq \int_0^t \|G_1(S)x_s\|^2\,ds \leqq c\int_0^t \|x_s\|^2\,ds \leqq \mu^2 c c_3 r_t,$$

which verifies (23) with $\varepsilon = \mu^2 c c_3$. $\square$

Example 2. Consider the following system:

$$A(S)y_t = SB(S)u_t + C(S)v_t + S\xi_t(y, u).$$

When the last term is identically zero, this model becomes the continuous-time analogue of an ARMAX model (see, e.g., [9]-[12]).

It is clear that Assumption 1 is verified if the nonlinear part $\xi_t(y, u)$ is one of the following forms:

$$\xi_t(y, u) = \varepsilon_1 y_t \sin(t) + \varepsilon_2 u_t \cos(t),$$

$$\xi_t(y, u) = \varepsilon_1 \sin(y_t) + \varepsilon_2 \sin(u_t), \qquad \varepsilon_i \in [0, \varepsilon], \quad i = 1, 2,$$

and so on.

3. Robustness of parameter estimation. We now show that the estimation error is proportional to the constant ε defined in (23) if the input-output data is persistently exciting.

THEOREM 1. *If Assumptions* 1 *and* 2 *are satisfied, then*

$$\limsup_{t\to\infty} \|\theta_t - \theta\| \leqq \alpha k \varepsilon, \quad a.s.$$

where $\alpha \in (0, \infty)$ *is a constant,* ε *is defined in* (23), *and*

$$k = \limsup_{t\to\infty} r_t / \lambda_{\min}(t) < \infty$$

where $\lambda_{\min}(t)$ *denotes the minimum eigenvalue of* P_t^{-1}.

For the proof of this theorem we need the following lemmas.

LEMMA 2. *Under the conditions of Theorem* 1, *there is a constant* $k_0 > 0$ *such that*

$$\operatorname{tr} \tilde{\theta}_t^\tau P_t^{-1} \tilde{\theta}_t \leqq O(1) + O\left(\left\{\int_0^t \|g_s\|^2\, ds\right\}^{(1/2)+\eta}\right) + O(\log r_t)$$
$$+ 2\left\{-\left(k_0 - \frac{c}{2}\right)\int_0^t \|g_s\|^2\, ds + \frac{1}{2c}\int_0^t \|D(S)C^{-1}(S)\dot{\eta}_s\|^2\, ds\right\},$$
$$\forall \eta > 0, \quad c > 0,$$

where $\tilde{\theta}_t = \theta - \theta_t$, $g_t = \tilde{\theta}_t^\tau \phi_t$.

Proof. By (7) and (16) it is easy to see that

$$\begin{aligned} dy_t &= \theta^\tau \phi_t^0\, dt + dv_t + \dot{\eta}_t\, dt \\ &= \theta^\tau \tilde{\phi}_t\, dt + \theta^\tau \phi_t\, dt + dv_t + \dot{\eta}_t\, dt \end{aligned}$$

and hence

$$\begin{aligned} \theta^\tau \tilde{\phi}_t\, dt &= dy_t - \theta_t^\tau \phi_t\, dt + (\theta_t - \theta)^\tau \phi_t\, dt - dv_t - \dot{\eta}_t\, dt \\ &= d\hat{v}_t - \tilde{\theta}_t^\tau \phi_t\, dt - dv_t - \dot{\eta}_t\, dt \\ &= -d\tilde{v}_t - \tilde{\theta}_t^\tau \phi_t\, dt - \dot{\eta}_t\, dt \end{aligned} \tag{31}$$

or

$$C(S)\left(\frac{d\tilde{v}_t}{dt}\right) = -g_t - \dot{\eta}_t \quad \text{or} \quad \left(\frac{d\tilde{v}_t}{dt}\right) = -C^{-1}(S)(g_t + \dot{\eta}_t). \tag{32}$$

Let us now set

$$f_t = \left\{\frac{[C(S) - D(S)]}{S}\right\}\tilde{v}_t + \frac{g_t}{2}; \tag{33}$$

it then follows that

$$f_t = \left[D(S)C^{-1}(S) - \frac{I}{2}\right]g_t + \left\{\frac{[D(S)C^{-1}(S) - I]}{S}\right\}\eta_t. \tag{34}$$

From this and Assumption 2 there are constants $k_0 > 0$, $k_1 > 0$ such that

$$\int_0^t g_s^\tau \{f_s + [I - D(S)C^{-1}(S)]\dot{\eta}_s - k_0 g_s]\, ds + k_1 > 0. \tag{35}$$

From (10), (32), and (33) it follows that

$$\begin{aligned} d\tilde{\theta}_t &= -P_t \phi_t D(S)[dy_t^\tau - \phi_t^\tau \theta_t\, dt] \\ &= -P_t \phi_t D(S)[dv_t - d\tilde{v}_t]^\tau \\ &= -P_t \phi_t [dw_t - d\tilde{v}_t - D_1 \tilde{v}_t\, dt - \cdots - D_r S^{r-1}\tilde{v}_t\, dt]^\tau \\ &= -P_t \phi_t \left[g_t\, dt + \dot{\eta}_t\, dt + \frac{C(S) - D(S)}{S}\,\tilde{v}_t\, dt + dw_t\right]^\tau \\ &= -P_t \phi_t \left(f_t\, dt + \frac{1}{2} g_t\, dt + \dot{\eta}_t\, dt + dw_t\right)^\tau. \end{aligned} \tag{36}$$

Applying Ito's formula, we obtain

$$\begin{aligned} d[\operatorname{tr} \tilde{\theta}_t^\tau P_t^{-1} \tilde{\theta}_t] &= -2g_t^\tau [f_t\,dt + \dot{\eta}_t\,dt + dw_t] + \phi_t^\tau P_t \phi_t\,dt \\ &= -2g_t^\tau \{f_t + [I - D(S)C^{-1}(S)]\dot{\eta}_t - k_0 g_t\}\,dt \\ &\quad + 2g_t^\tau [1 - D(S)C^{-1}(S)]\dot{\eta}_t\,dt - 2k_0 \|g_t\|^2\,dt \\ &\quad - 2g_t^\tau \dot{\eta}_t\,dt - 2g_t^\tau\,dw_t + \phi_t^\tau P_t \phi_t\,dt; \end{aligned}$$

then by (35)

$$\begin{aligned} 0 &\leqq \operatorname{tr} \tilde{\theta}_t^\tau P_t^{-1} \tilde{\theta}_t \\ &\leqq \operatorname{tr} \tilde{\theta}_0^\tau P_0^{-1} \tilde{\theta}_0 + \int_0^t \phi_s^\tau P_s \phi_s\,ds + 2k_1 \\ &\quad + 2\left\{-k_0 \int_0^t \|g_s\|^2\,ds - \int_0^t g_s^\tau D(S)C^{-1}(S)\dot{\eta}_s\,ds - \int_0^t g_s^\tau\,dw_s\right\}. \end{aligned} \tag{37}$$

Noting the following elementary facts:

$$2\int_0^t a_s^\tau b_s\,ds \leqq c\int_0^t \|a_s\|^2\,ds + c^{-1}\int_0^t \|b_s\|^2\,ds \quad \forall c > 0,$$

$$\begin{aligned} \int_0^t \phi_s^\tau P_s \phi_s\,ds &= \int_0^t \operatorname{tr}[P_s \phi_s \phi_s^\tau\,ds] = \int_0^t \operatorname{tr}[P_s\,dP_s^{-1}] \\ &= \int_0^t \frac{d(\det P_s^{-1})}{\det P_s^{-1}} = O(\log r_t), \end{aligned}$$

and applying the following estimate for the Ito integral (see, e.g., [16, Lemma 4]):

$$\int_0^t x_s^\tau\,dw_s = O(1) + o\left(\left\{\int_0^t \|x_s\|^2\,ds\right\}^{1/2+\eta}\right) \quad \text{a.s. } \forall \eta > 0, \tag{38}$$

for any predictable process (x_t, F_t), we can easily conclude the lemma from (37). □

LEMMA 3. *If F_d defined by* (21) *is stable, then*

$$\frac{1}{t}\int_0^t V_s V_s^\tau\,ds \to R \quad \text{a.s. as } t \to \infty \tag{39}$$

where V_t is defined in (19) *and*

$$R \triangleq \int_0^\infty \exp\{F_d \lambda\} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \exp\{F_d^\tau \lambda\}\,d\lambda.$$

Proof. Since F_d is stable, there exists a positive-definite matrix $P > 0$ such that

$$PF_d + F_d^\tau P = -I.$$

By this and the Ito formula we see from (22) that

$$\begin{aligned} d[V_t^\tau P V_t] &= V_t^\tau (PF_d + F_d^\tau P) V_t\,dt + \operatorname{tr}\left\{\begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} [I, 0 \cdots 0] P\,dt + 2V_t^\tau P \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} dw_t \right. \\ &= -\|V_t\|^2\,dt + \operatorname{tr}\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} P\,dt + 2V_t^\tau P \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} dw_t. \end{aligned}$$

So it follows by applying (38) that

$$V_t^\tau P V_t + \int_0^t \|V_s\|^2\,ds = \operatorname{tr}\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Pt + o\left(\left\{\int_0^t \|V_s\|^2\,ds\right\}^{1/2+\eta}\right). \tag{40}$$

Consequently, we conclude that

$$\int_0^t \|V_s\|^2\,ds = O(t), \quad \text{a.s.} \tag{41}$$

Again, by the Ito formula we get

$$V_t V_t^\tau = \left(\int_0^t V_s V_s^\tau\,ds\right) F_d^\tau + F_d\left(\int_0^t V_s V_s^\tau\,ds\right) + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} t + \int_0^t \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} dw_s V_s^\tau + \int_0^t V_s\,dw_s^\tau [I, 0 \cdots 0]$$

and hence

$$\begin{aligned} &\int_0^t V_s V_s^\tau\,ds \\ &= \int_0^t \exp[F_d(t-z)] \int_0^z \left\{ \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} dw_s V_s^\tau + V_s\,dw_s^\tau [I, 0 \cdots 0] \right\} \exp[F_d^\tau (t-z)]\,dz \\ &\quad + \int_0^t \exp[F_d(t-z)] \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} z \exp[F_d^\tau (t-z)]\,dz. \end{aligned} \tag{42}$$

We now consider the first term on the right-hand side (42). By (38), (41), and the stability of F_d, it is easy to see that there is a constant $\rho > 0$ such that

$$\begin{aligned} &\left\| \int_0^t \exp[F_d(t-z)] \int_0^t \left\{ \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} dw_s V_s^\tau + V_s\,dw_s^\tau [I, 0 \cdots 0] \right\} \exp[F_d^\tau (t-z)]\,dz \right\| \\ &= O\left(\int_0^t \exp[-2\rho(t-z)] \left\{ \int_0^z \|V_s\|^2\,ds \right\}^{1/2+\eta} dz \right) + O(1) \\ &= O\left(\int_0^t \exp[-2\rho(t-z)] z^{1/2+\eta}\,dz \right) = O(t^{1/2+\eta}) \quad \forall \eta > 0. \end{aligned}$$

Hence the lemma follows immediately from this and (42).

Proof of Theorem 1. Since $D(S)C^{-1}(S)$ is strictly positive real, $C(S)$ is stable. Then by Lemma 1 and Assumption 1, it follows that

$$\int_0^t \|D(S)C^{-1}(S)\dot{\eta}_s\|^2\,ds \leqq \varepsilon c_0 r_t \quad \text{for some } c_0 > 0.$$

Taking $c < 2k_0$ in Lemma 2, we see that

$$0 \leqq \operatorname{tr} \tilde{\theta}_t^\tau P_t^{-1} \tilde{\theta}_t \leqq O(1) - \left(k_0 - \frac{c}{2}\right) \int_0^t \|g_s\|^2\,ds + O(\varepsilon r_t) + O(\log r_t). \tag{43}$$

Then for sufficiently large t

$$\begin{aligned}
\operatorname{tr} \tilde{\theta}_t^\tau \tilde{\theta}_t &\leqq \frac{\operatorname{tr}(\tilde{\theta}_t^\tau P_t^{-1} \tilde{\theta}_t)}{\lambda_{\min}(t)} \\
&\leqq \frac{1}{\lambda_{\min}(t)} \left\{ O(1) - \left(k_0 - \frac{c}{2}\right) \int_0^t \|g_s\|^2 \, ds + O(\varepsilon r_t) + O(\log r_t) \right\} \qquad (44) \\
&\leqq O\left(\frac{\log r_t}{r_t}\right) - \frac{k}{r_t}\left(k_0 - \frac{c}{2}\right) \int_0^t \|g_s\|^2 \, ds + O(\varepsilon k).
\end{aligned}$$

Since $c < 2k_0$, the desired result will follow if we can show that $r_t \to \infty$, as $t \to \infty$. We prove this as follows.

From (32) it follows that

$$\tilde{V}_t = -\int_0^t \exp\{F_c(t-s)\}[g_s + \dot{\eta}_s] \, ds.$$

Then by (43) and Assumption 1, we have for some $\rho > 0$ and $c_1 > 0$

$$\begin{aligned}
\int_0^t \|\tilde{V}_z\|^2 \, dz &\leqq \int_0^t \left\| \int_0^z \exp\{F_c(z-s)\}[g_s + \dot{\eta}_s] \, ds \right\|^2 dz \\
&\leqq (c_1)^2 \int_0^t \left\{ \int_0^z \exp[-\rho(z-s)][\|g_s\| + \|\dot{\eta}_s\|] \, ds \right\}^2 dz \\
&\leqq 2(c_1)^2 \int_0^t \int_0^z \exp[-\rho(z-s)] \, ds \int_0^z \exp[-\rho(z-s)][\|g_s\|^2 + \|\dot{\eta}_s\|^2] \, ds \, dz \\
&\leqq 2\rho^{-1}(c_1)^2 \int_0^t \int_z^t \exp[-\rho(z-s)] \, dz \, [\|g_s\|^2 + \|\dot{\eta}_s\|^2] \, ds \qquad (45) \\
&\leqq 2\rho^{-2}(c_1)^2 \int_0^t [\|g_s\|^2 + \|\dot{\eta}_s\|^2] \, ds \\
&\leqq 2\rho^{-2}(c_1)^2 \{O(\log r_t) + O(\varepsilon r_t) + \varepsilon r_t\} \\
&= O(\log r_t) + O(\varepsilon r_t).
\end{aligned}$$

Assume the converse were true, i.e., r_t was bounded in t; then from (45) it would follow that $\int_0^t \|\tilde{V}_z\|^2 \, dz$ would be bounded. But by (20) and (25) it is clear that

$$r_t \geqq \int_0^t \|\hat{V}_z\|^2 \, dz = \int_0^t \|V_z\|^2 \, dz + \int_0^t \|\tilde{V}_z\|^2 \, dz - 2\int_0^t V_z^\tau \tilde{V}_z \, dz.$$

From this and the boundedness of r_t and $\int_0^t \|\tilde{V}_z\|^2 \, dz$, it follows that

$$\int_0^t \|V_z\|^2 \, dz \text{ is bounded.}$$

This contradicts Lemma 3. Hence $r_t \to \infty$, a.s., and Theorem 1 holds. □

Remark 1. If in (7) the unmodeled dynamics $\{\eta_t\}$ are identically zero, then we may take ε as zero in Assumption 1. In this case, it follows from (44) that

$$\|\tilde{\theta}_t\|^2 = O\left(\frac{\log r_t}{\lambda_{\min}(t)}\right) \quad \text{a.s.}$$

This result is the continuous-time version of that obtained in the discrete-time case (see, e.g., [13]-[14]). See also [12] for related results.

4. Robustness of adaptive tracking. Let $\{u_t^*\}$ be a bounded deterministic and differentiable reference signal with $u_0^* = 0$. Our objective here is to design the adaptive control u_t, so that the output $\{y_t\}$ tracks the output of the following reference model:

$$E(S)y_t^* = u_t^*$$

where $E(S) = I + E_1 S + \cdots + E_p S^p$ is a stable matrix polynomial.

Similar to (18), we set

$$Y_t^* = [y_t^* \cdots S^{p-1} y_t^*]^\tau.$$

By a representation similar to (22), it is easy to see that $\{Y_t^*\}$ is a bounded sequence.

From now on, we assume that the upper bound for the order of the polynomial $A(S)$ is equal to that of $C(S)$, i.e., $p = r$.

Similar to the discrete-time case, we need the following standard minimum phase condition.

Assumption 3. $B(S)$ is stable.

Let us define the adaptive control u_t via the following equation:

$$\theta_t^\tau \phi_t = \frac{dy_t^*}{dt}. \tag{46}$$

This together with (1), (10), and (11) form a system of nonlinear stochastic differential equations, for which the existence and uniqueness of the strong solution is assumed.

THEOREM 2. *Consider the system* (1)-(6) *with* $p = r$, *and the estimation algorithm* (10)-(11). *If Assumptions* 1-3 *hold, and the control law is defined from* (46), *then there exists* $\varepsilon_1 > 0$ *such that whenever* ε *in* (23) *lies in the interval* $[0, \varepsilon_1)$, *the following properties hold*:

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T (\| Y_t \|^2 + \| U_t \|^2)\, dt < \infty \quad a.s. \tag{47}$$

and

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T \| Y_t - Y_t^* \|^2\, dt = \operatorname{tr} R + \delta \quad a.s. \tag{48}$$

where $|\delta| = O(\varepsilon^{1/2})$, *and* R *is defined in Lemma* 3.

Proof. From (13) and (46) it is easy to see that

$$y_t - y_t^* = \hat{v}_t, \qquad Y_t^* - Y_t = \tilde{V}_t - V_t. \tag{49}$$

Note that

$$r_T = \int_0^T (\| Y_t \|^2 + \| U_t \|^2 + \| \hat{V}_t \|^2)\, dt + e. \tag{50}$$

We have by (39), (45), and (49) that

$$\int_0^T \| Y_t \|^2\, dt = O(T) + \varepsilon c_3 r_T + O(\log r_T), \qquad c_3 > 0. \tag{51}$$

From this, (7) and the stability of $B(S)$, we have

$$\int_0^T \| U_t \|^2\, dt = O(T) + \varepsilon c_4 r_T + O(\log r_T), \qquad c_4 > 0. \tag{52}$$

Note also that by (45) and Lemma 3, we have

$$\int_0^T \|\hat{V}_t\|^2\,dt \leqq O(T)+2\int_0^T \|\tilde{V}_t\|^2\,dt \leqq O(T)+\varepsilon c_5 r_T + O(\log r_T).$$

Hence, combining (50)-(52), we have

$$r_t \leqq O(t)+\varepsilon c_6 r_t + O(\log r_t), \qquad c_6>0,$$

which yields

$$\limsup_{t\to\infty}\frac{r_t}{t}<\infty \quad \text{for any } \varepsilon\in[0,\varepsilon_1)$$

with $\varepsilon_1 = 1/c_6$. Thus (47) is true.

We now proceed to prove (48). From (45) we have for any $\varepsilon\in[0,\varepsilon_1)$,

$$\frac{1}{T}\int_0^T \|\tilde{V}_t\|^2\,dt = O\left(\frac{\log T}{T}\right)+O(\varepsilon); \tag{53}$$

then by (49)

$$\begin{aligned}
&\frac{1}{T}\int_0^T (Y_t-Y_t^*)(Y_t-Y_t^*)^\tau\,dt\\
&\quad=\frac{1}{T}\int_0^T (V_t-\tilde{V}_t)(V_t-\tilde{V}_t)^\tau\,dt\\
&\quad=\frac{1}{T}\int_0^T V_tV_t^\tau\,dt+\frac{1}{T}\int_0^T \tilde{V}_t\tilde{V}_t^\tau\,dt-\frac{1}{T}\int_0^T (V_t\tilde{V}_t^\tau+\tilde{V}_tV_t^\tau)\,dt,\\
&\quad=R+\left[\frac{1}{T}\int_0^T V_tV_t^\tau\,dt-R\right]+O\left(\frac{\log T}{T}\right)+O(\varepsilon)+O\left(\left\{\frac{\log T}{T}+\varepsilon\right\}^{1/2}\right).
\end{aligned} \tag{54}$$

Hence (48) is also true. □

Remark 2. If the initial value of the reference signal is not zero, i.e., $u_0^*\neq 0$, then we may replace (46) by

$$\theta_t^\tau\phi_t=\frac{dz_t^*}{dt},$$

where $z_t^*=E^{-1}(S)\{u_t^*-\exp(-t^2)u_0^*\}$. In this case, Theorem 2 is true for $\{z_t^*\}$, which approximates $\{y_t^*\}$ exponentially.

5. Asymptotic behavior of adaptive tracking. In this section we assume $\eta_t=0$ in (7). For this ideal case we give the convergence rate for the adaptive tracking errors. It is worth noting that the corresponding discrete-time results have not yet been established (see, e.g., [17], for related discussions).

LEMMA 4. *Let $\{x_t\}$ be any measurable process adapted to $\{F_t\}$, satisfying*

$$\limsup_{T\to\infty}\frac{1}{T}\int_0^T \|x_t\|^2\,dt\leqq k_1<\infty \quad a.s.$$

for some constant k_1. Then

$$\limsup_{T\to\infty}\frac{1}{(T\log\log T)^{1/2}}\left\|\int_0^T x_t\,dw_t^\tau\right\|<\infty \quad a.s. \tag{55}$$

Proof. Without loss of generality, we assume that x_t and w_t are scalars. Taking a constant k_2 so large that

$$(k_2)^2 - k_2 + 2(1-k_2)k_1 > 0, \qquad k_2 > 1,$$

we have

$$\begin{aligned}\int_0^T (k_2 + x_t)^2\, dt &\geqq \int_0^T (x_t)^2\, dt + (k_2)^2 T - 2k_2 \int_0^T |x_t|\, dt \\ &\geqq \int_0^T (x_t)^2\, dt + (k_2)^2 T - k_2 \left[T + \int_0^T (x_t)^2\, dt \right] \\ &\geqq k_2(k_2-1)T + 2(1-k_2)k_1 T.\end{aligned}$$

Consequently,

$$\int_0^\infty (k_2 + x_t)^2\, dt = \infty \quad \text{a.s.}$$

Now, define the following stopping time:

$$\tau(t) = \inf\left\{ s: \int_0^s (k_2 + x_z)^2\, dz = t \right\}.$$

It is known that

$$\int_0^{\tau(t)} (k_2 + x_s)\, dw_s$$

is a Brownian motion (see, e.g., [18, Thm. 4.5]). Then by the law of the iterated logarithm for Wiener processes, we have

$$\frac{1}{(t \log\log t)^{1/2}} \left| \int_0^{\tau(t)} (k_2 + x_s)\, dw_s \right| = O(1) \quad \text{a.s.} \tag{56}$$

Denoting

$$a(t) = \int_0^t (k_2 + x_z)^2\, dz, \tag{57}$$

it is evident that $a(\tau(t)) = t$. Then (56) and (57) imply

$$\frac{1}{[a(T) \log\log a(T)]^{1/2}} \left| \int_0^T (k_2 + x_s)\, dw_s \right| = O(1) \quad \text{a.s.}$$

as $T \to \infty$. From this and the fact that $a(T)/T = O(1)$, it follows that

$$\frac{1}{[T \log\log T]^{1/2}} \left| \int_0^T (k_2 + x_s)\, dw_s \right| = O(1) \quad \text{a.s.}$$

and hence

$$\begin{aligned}&\frac{1}{[T \log\log T]^{1/2}} \left| \int_0^T x_s\, dw_s \right| \\ &\quad \leqq \frac{1}{[T \log\log T]^{1/2}} \left\{ k_2 |w_T| + \left| \int_0^T (k_2 + x_s)\, dw_s \right| \right\} \\ &\quad = O(1) \quad \text{a.s. as } T \to \infty,\end{aligned}$$

completing the proof. □

We are now in a position to prove the following main result of this section.

THEOREM 3. *Consider the system described by* (7) *with* $\eta_t = 0$ *and* $p = r$, *and estimation algorithm* (10)-(11). *If Assumptions* 2 *and* 3 *are satisfied, and if the adaptive control is defined from* (46), *then*

$$\|R_T - R\|^2 = O\left(\frac{\log T}{T}\right) \quad a.s.\ as\ T \to \infty, \tag{58}$$

where R is given in Lemma 3 *and*

$$R_T = \frac{1}{T}\int_0^T (Y_t - Y_t^*)(Y_t - Y_t^*)^\tau\, dt.$$

Proof. We first consider the convergence rate of $1/T\int_0^T V_t V_t^\tau\, dt$.

From (40) it is clear that

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T \|V_s\|^2\, ds \leqq 2\,\mathrm{tr}\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} P.$$

Then Lemma 4 implies

$$\limsup_{T\to\infty} \frac{1}{(T\log\log T)^{1/2}} \left\| \int_0^T V_t\, dw_t^\tau \right\| < \infty \quad \text{a.s.},$$

and hence

$$\left\| \int_0^t \exp\left[F_d(t-z)\right] \int_0^z \left\{ \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} dw_s\, V_s^\tau + V_s\, dw_s^\tau [I, 0 \cdots 0] \right\} \exp\left[F_d^\tau (t-z)\right] dz \right\|$$

$$= O(\{t\log\log t\}^{1/2}) \quad \text{a.s.}$$

Consequently, it follows from (42) that

$$\left\| \frac{1}{T}\int_0^T V_s V_s^\tau\, ds - \int_0^\infty \exp\{F_d s\} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \exp\{F_d^\tau s\}\, ds \right\| = O\left(\left\{\frac{\log\log T}{T}\right\}^{1/2}\right). \tag{59}$$

Setting $\varepsilon = 0$ in (53) and (54), and using (59), we see that

$$R_T = \int_0^\infty \exp\{F_d s\} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \exp\{F_d^\tau s\}\, ds + O\left(\left\{\frac{\log\log T}{T}\right\}^{1/2}\right) + O\left(\frac{\log T}{T}\right) + O\left(\left\{\frac{\log T}{T}\right\}^{1/2}\right)$$

$$= \int_0^\infty \exp\{F_d s\} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \exp\{F_d^\tau s\}\, ds + O\left(\left\{\frac{\log T}{T}\right\}^{1/2}\right),$$

which verifies (58). Hence the proof is complete. □

REFERENCES

[1] B. EGARDT, *Stability analysis of adaptive control systems with disturbances*, Proc. JACC, San Francisco, CA, 1980.

[2] C. E. ROHRS, L. VALAVANI, M. ATHANS, AND C. STEIN, *Robustness of adaptive control algorithm in the presence of unmodeled dynamics*, Preprints of 21st IEEE Conference on Decision and Control, Orlando, FL, 1982.

[3] B. D. RIEDLE AND P. V. KOKOTOVIC, *Disturbance instabilities in an adaptive system*, IEEE Trans. Automat. Control, 29 (1984), pp. 822-824.

[4] R. ORTEGA, L. PRALY, AND I.D. LANDAU, *Robustness of discrete-time direct adaptive controllers*, IEEE Trans. Automat. Control, 30 (1985).

[5] G. KREISSELMEIER AND B. D. O. ANDERSON, *Robust model reference adaptive control*, IEEE Trans. Automat. Control, 31 (1986).

[6] P. A. IOANNOU AND K. TSAKLIS, *A robust direct adaptive controller*, IEEE Trans. Automat. Control, 31 (1986).

[7] H. F. CHEN AND L. GUO, *Robustness analysis of identification and adaptive control for stochastic systems*, Systems Control Lett., 9 (1987), pp. 131-140.

[8] ———, *A robust stochastic adaptive controller*, IEEE Trans. Automat. Control, 33 (1988), pp. 1035-1043.

[9] H. F. CHEN, *Quasi-least-squares identification and its strong consistency*, Internat. J. Control, 34 (1981), pp. 921-936.

[10] J. H. VAN SCHUPPEN, *Convergence results for continuous time adaptive stochastic filtering algorithms*, J. Math. Anal. Appl., 96 (1983), pp. 209-225.

[11] H. F. CHEN, *Recursive Estimation and Control for Stochastic Systems*, John Wiley, New York, 1985.

[12] H. F. CHEN AND J. B. MOORE, *Convergence rate of continuous time stochastic* ELS *parameter estimation*, IEEE Trans. Automat. Control, 32 (1987), pp. 267-269.

[13] T. L. LAI AND C. Z. WEI, *Extended least squares and their application to adaptive control and prediction in linear systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 898-906.

[14] H. F. CHEN AND L. GUO, *Convergence rate of least squares identification and adaptive control for stochastic systems*, Internat. J. Control, 44 (1986), pp. 1459-1476.

[15] R. S. LIPSTER AND A. N. SHIRYAYEV, *Statistics of Random Processes*, I. *General Theory*, Springer-Verlag, New York, 1977.

[16] N. CHRISTOPEIT, *Quasi-least-squares estimation in semimartingale regression models*, Stochastics, 16 (1986), pp. 255-278.

[17] P. R. KUMAR, *Convergence of adaptive control schemes using least-squares parameter estimates*, 1989, submitted.

[18] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, Vol. 1, Academic Press, New York, 1975.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 528-539, May 1990

003

GLOBAL OPTIMIZATION OVER UNBOUNDED DOMAINS*

H. RATSCHEK† AND R. L. VOLLER†

Abstract. Almost all methods for solving the global optimization problem need the assumption that a parallelepiped containing the solution points is known. The boundedness is necessary both for the numerical computation as well as for guaranteeing the convergence properties. In this paper a technique is described that drops this restriction so that the unconstrained problem, in the literal sense of the terms can be solved. The technique is based on the branch-and-bound method and on infinite-interval arithmetic, it is simple to apply, and very robust as examples show.

Key words. global optimization, unconstrained optimization, nonlinear optimization, optimization over unbounded domains

AMS(MOS) subject classifications. 90C30, 65K05, 65K10, 65G10

1. Introduction. Almost all of the well-known methods for solving the global unconstrained optimization problem involve a parallelepiped as bounded subdomain $X \subseteq R^m$ for the function to which the method is applied. Therefore, X must be known a priori or must be determined by means of an analysis of the problem. (A notable exception is [12].)

The technique we provide is destined to solve the global minimization problem for a continuous function $f: R^m \to R$. The search for the global minimizers is accomplished in the whole domain R^m. Bounds of the global minimum f^* are generated, and one or several boxes of prescribed size that include all global minimizers are produced. It can be detected when f has no global minimum at all, and further, whether or not f is bounded from below. If the technique is applied on a computer, the sharpness of these detections is limited by the finite number representation of computers. Although weakened, the detections remain logically valid in this case such that the user can trust them.

The method that we provide to cover the whole space R^m when looking for global minimizers is best demonstrated by applying it to Hansen's Algorithm [3]. Similar algorithms, such as those of Skelboe [17], Moore [10], Asaithambi, Shen, and Moore [1], and Ichida-Fujii [4] may also be used.

Hansen's Algorithm is very sophisticated, hence we discuss a simplified prototype version in § 2. In §§ 3 and 4, the optimization problem and the prototype algorithm are extended and admitted to functions that are defined on R^m. For this reason, a compactification of the space R^m is introduced, $R^m := (\bar{R})^m$ where $\bar{R} := R \cup \{\infty, -\infty\}$. The advantages of this compactification are threefold: (i) The investigation of the convergence properties can make use of compactness principles. (ii) The interpretation in R^m of the results obtained in R^m is straightforward and there is no need to distinguish between the bounded and the unbounded case. (iii) The step from the exact execution of the algorithm in R^m to its numerical execution on a computer is small because the latter operates in $[-L, L]^m$ where L denotes the largest representable number of the machine under consideration. R^m and $[-L, L]^m$ are topologically very similar.

In § 5, the monotonicity test [3], [10] is discussed. This is an extremely effective means for detecting global monotonicity of f in a box such that this box can be

* Received by the editors November 2, 1987; accepted for publication (in revised form) May 17, 1989.
† Mathematisches Institut der Universität Düsseldorf, Universitätsstrasse 1, D-4000 Düsseldorf 1, Federal Republic of Germany.

discarded from the search for minimizers. This test is carried over to the unbounded case and the usual assumption for f to be differentiable is weakened. In § 6 an arithmetic for unbounded noncompact intervals is introduced to obtain the inclusion functions required for the extended algorithm. As a consequence, Moore's principle of natural interval extension [9] can be recursively defined for programmable functions over unbounded domains. In § 7, the relationships between the numerical and the exact realization of the extended algorithm are discussed. In § 8, numerical examples show that the practical computation involves no difficulties at all.

Comparing §§ 3, 6, and 7, we are faced with three kinds of infinite intervals:

(i) *Compactified unbounded intervals* such as $[a, \infty] \subseteq \bar{R}$ (cf. § 3). They are needed both for the execution of the algorithm and for the discussion of its convergence properties. Since only topological arguments and no arithmetic are used for this discussion, no arithmetic need be defined for compactified unbounded intervals.

(ii) *Unbounded noncompactified intervals*, such as $[a, \infty) \subseteq R$ (cf. § 6). They occur when the bounds of f over unbounded subdomains are determined. Thus, an arithmetic for such intervals is defined.

(iii) Both kinds of intervals mentioned in (i) and (ii) must be simulated by *machine intervals* when computing on a machine (cf. § 7).

2. The algorithm over bounded domains. We start the discussion with a prototype version of Hansen's Algorithm [3] that is intended to solve optimization problems over parallelepipeds. Hansen's Algorithm differs from the prototype in that it contains many excellent further techniques that speed up the calculation, but do not change the convergence order and the solution set. To have a straightforward discussion we drop these techniques.

Let R be the set of reals, and I be the set of real compact intervals. Right parallelepipeds such as $Y = Y_1 \times \cdots \times Y_m \in I^m$ are called *boxes*. If $D \subseteq R^m$ then $I(D)$ denotes the set of all boxes $Y \subseteq D$. The width of an interval is denoted by $w([a, b]) = b - a$. The *width* of a box Y is defined by $w(Y) = \max_{i=1,\cdots,m} w(Y_i)$. Let $f: D \to R$. The *range* of f over $Y \subseteq D$ is denoted by $\square f(Y) = \{f(y): y \in Y\}$. An interval function $F: I(D) \to I$ is called an *inclusion function* for f if $\square f(Y) \subseteq F(Y)$ for any $Y \in I(D)$. If $A = [a, b]$ is an interval, the lower and upper boundaries of A are denoted by $\operatorname{lb} A = a$ and $\operatorname{ub} A = b$, respectively. The midpoint is denoted by $\operatorname{mid} A = (a + b)/2$, and the midpoint of a box $Y \in I^m$ by $\operatorname{mid} Y = (\operatorname{mid} Y_i)_{i=1}^m$.

The following prototype version of Hansen's Algorithm aims to determine the global minimum and the global minimizers of a function $f: X \to R$ over a bounded box $X \in I^m$ when an inclusion function F for f is given.

Algorithm 1.

(1) Set $Y := X$.
(2) Calculate $F(Y)$, $f(c)$ where $c = \operatorname{mid} Y$.
(3) Set $y := \operatorname{lb} F(Y)$.
(4) Initialize list $L := \{(Y, y)\}$. Set $\tilde{f} := f(c)$.
(5) Choose a coordinate direction k parallel to which Y has an edge of maximum length, that is, $k \in \{i: w(Y_i) = w(Y)\}$.
(6) Bisect Y in direction k getting boxes V_1, V_2 such that $Y = V_1 \cup V_2$.
(7) Calculate $F(V_1)$, $F(V_2)$.
(8) Set $v_i := \min F(V_i)$ for $i = 1, 2$.
(9) Enter the pairs (V_1, v_1), (V_2, v_2) at the end of the list.
(10) (Optional) If f has a generalized gradient in each point of $V_i \cap R^m$ then apply the monotonicity test to V_i for $i = 1, 2$.

(11) Choose a pair $(\tilde{Y}, \tilde{y})$ of the list that satisfies $\tilde{y} \leqq z$ for all pairs (Z, z) of the list.
(12) Discard all pairs (Z, z) from the list that satisfy $\tilde{f} < z$ (*midpoint test*).
(13) If termination criteria hold go to (16).
(14) Denote the first pair of the list by (Y, y). Set $c := \operatorname{mid}(Y)$ and $\tilde{f} := \min(\tilde{f}, f(c))$.
(15) Go to (5).
(16) End

The *monotonicity test* (cf. step (10)) is a means to detect strict monotonicity of f in a box that can then be deleted. The test is a device for accelerating the computation rather than a substantial part of the algorithm since its solution set is independent of the use of the test. The situation, however, changes when unbounded domains are admitted such that the test is highly recommended. The detailed discussion is referred to in §§ 5 and 7.

The midpoint test (cf. step (12)) may operate any $c \in Y$ instead of mid Y. Also ub $F(Y)$ can be used instead of $f(c)$.

A thorough investigation of *termination criteria* can be found in [3] and [11].

Algorithm 1 produces an infinite sequence of lists (L_n) if termination by step (13) is dropped. Furthermore, we get sequences $(\tilde{y}_n)$ and (f_n) according to steps (11) and (14). (We write f_n instead of $\tilde{f}_n$.) Let U_n be the union of all boxes occurring in L_n. It is obvious that (U_n) is a nested sequence. The *convergence properties* of Algorithm 1 are discussed in [11] and [16]. Hints for the construction of inclusion functions and the related order are given in [15].

3. The algorithm over unbounded domains. Algorithm 1 shows very pleasant convergence properties [11], [16]. They depend on the compactness of X. Hence, we will first provide a compactification of R^m, say $\overline{R^m}$. Then both, Algorithm 1 as well as the convergence properties are extended to the compactified unbounded case. For this reason, the objective function f and its inclusion function F must be extended to $\overline{R^m}$. The results gained in $\overline{R^m}$ are then converted into the originally wanted results in R^m. The compactification could be avoided but simplifies matters considerably.

To apply Algorithm 1 to $\overline{R^m}$, the midpoint and the width of boxes in $\overline{R^m}$ must be defined, and further, the given function f and its inclusion function F must be extended. This requires some notation.

Let $\bar{R} := R \cup \{-\infty, \infty\} = [-\infty, \infty]$ be the two-point compactification of R, and let $\overline{R^m} := \bar{R}^m$ be the m-fold topological product of $\bar{R}$. If $A \subseteq R^m$, we denote the compact hull of A with respect to this compactification by $\bar{A}$.

Let I_∞ be the set of all closed (but not necessarily bounded) intervals of R. Thus, the intervals $[a, b]$, $[a, \infty)$, $(-\infty, b]$, and $(-\infty, \infty) = R$ belong to I_∞ where $a, b \in R$. Let $\bar{I}$ be the set of all compact intervals of $\bar{R}$. Thus, $[a, b]$, $[a, \infty]$, $[-\infty, b]$, $[-\infty, \infty] = \bar{R}$, $\infty = [\infty, \infty]$, or $-\infty = [-\infty, -\infty]$ belong to $\bar{I}$ where $a, b \in R$.

Let $A \subseteq \overline{R^m}$; then $I_\infty(A) := \{Y \in I_\infty^m : Y \subseteq A\}$ and $\bar{I}(A) := \{Y \in \bar{I}^m : Y \subseteq A\}$. We note that $I_\infty^m = I_\infty(R^m) = I_\infty(\overline{R^m})$ and $\bar{I}^m = \bar{I}(\overline{R^m})$. Furthermore, $I(A) := \{Y \in I^m : Y \subseteq A\}$.

A width for unbounded boxes may be defined in a variety of ways resulting in formulas of greater or lesser complexity. We do not expect our formula to be either elegant or of theoretical interest (as the chordal-distance on the Riemann sphere), however, it must be appropriate for our purposes, which are the following: (i) If a nested box sequence (Y_n), $Y_n \in \bar{I}^m$ tends to a point of $\overline{R^m}$, the widths of the boxes must tend to zero. (ii) The width must control the bisection process (cf. steps (4) and (5) of Algorithm 1) to generate box sequences contracting to one point; (iii) These

properties of the width should be maintained at the numerical implementation, i.e., the width of the largest machine-infinite interval should be more or less comparable with the width of the smallest machine-finite interval (cf. § 7).

Our width concept depends on a global parameter λ that the user or programmer may choose, such that the global minimizers are suspected to lie in the box $[-\lambda, \lambda]^m$.

If the choice of the user is wrong, the program is still correct but slower. The choice of λ influences the bisection process, and areas outside of this box are treated as nearly infinite areas. For this reason, the width of intervals lying outside of $[-\lambda, \lambda]$ is adapted: If $-\infty < a \leqq b < \infty$, then

$$w([a, b]) := \lambda^2(b-a)/(ab) \quad \text{if } a \geqq \lambda \quad \text{or } b \leqq -\lambda.$$

The width of a box $Y = Y_1 \times \cdots \times Y_m \in \bar{I}^m$ is then defined as follows. Let $0 < \lambda < \infty$ and $a \in R$. Then

$$w([a, \infty]) := \begin{cases} \lambda^2/a & \text{if } a \geqq 10^{-10}, \\ 10^{10} \max(1, \lambda^2) & \text{otherwise,} \end{cases}$$

$$w([-\infty, a]) := w([-a, \infty]),$$

$$w([-\infty, \infty]) := 10^{11} \max(1, \lambda^2), \; w(\pm[\infty, \infty]) = 0,$$

$$w(Y) := \max_{i=1,\cdots,m} w(Y_i).$$

In order not to dissect $[-\lambda, \lambda]^m$ too early, the midpoint of unbounded boxes will also be made dependent on λ. Let $a \in \bar{R}$ and $Y \in \bar{I}^m$. Then we set

$$\operatorname{mid}[a, \infty] := \begin{cases} \lambda & \text{if } a < \lambda, \\ 2a & \text{if } \lambda \leqq a, \end{cases}$$

$$\operatorname{mid}[-\infty, a] := -\operatorname{mid}[-a, \infty] \quad \text{if } a < \infty,$$

$$\operatorname{mid} Y := (\operatorname{mid} Y_i)_{i=1}^m.$$

If, for example, $\lambda = 10$, then the interval $[20, \infty]$ (of width 5) is bisected into $[20, 40]$ and $[40, \infty]$ (both of width $\frac{5}{2}$).

Let $A \in I_\infty^m$ and $f: A \to R$ be given. We want to extend f to a function $f_0: \bar{A} \to \bar{R}$. Let $x \in \bar{A} \backslash A$, then we set

$$f_0(x) := \min \left\{ \lim_{n \to \infty} \inf f(x_n) : x_n \in A, \, x_n \to x \right\} \tag{1}$$

where the convergence of the sequences (x_n) to x is subject to the topology of $\overline{R^m}$. Note that f_0 need not be continuous, even when f is. For example, if $f(x) = e^x$, $x \in R$, then $f_0(-\infty) = 0$ and $f_0(\infty) = \infty$. If $f(x) = \sin x$, $x \in R$, then $f_0(-\infty) = f_0(\infty) = -1$.

Analogously, if $F: I(A) \to I$ is an inclusion function for f, we want to extend F to an inclusion function $F_0: \bar{I}(\bar{A}) \to \bar{I}$ for f_0, that is,

$$\square f_0(\bar{Y}) \subseteq F_0(\bar{Y}) \quad \text{for any } Y \in I_\infty(A). \tag{2}$$

We do not need inclusions of f over boxes $Z = Z_1 \times \cdots \times Z_m$ where any component Z_i is just ∞ or $-\infty$. Such cases are neither considered in (2) nor in the following definition (3), which simplifies matters. Thus we define F_0 by

$$F_0(\bar{Y}) := \overline{F(Y)} \quad \text{for } Y \in I_\infty(A). \tag{3}$$

We call F and also F_0 *nonwasteful* if, given any $Y \in I_\infty(A)$, a partition of A into bounded boxes exists, $A = \bigcup_{i \in J} B_i$, with $B_i \in I(A)$ and some index set J, such that

$$F(Y) \subseteq \overline{\bigcup_{i \in J} F(B_i)}. \tag{4}$$

Condition (4) is necessary for getting reasonable convergence properties. It is a very natural condition, since each programmer would automatically construct nonwasteful inclusions. For example, let $f(x) = x_1^2 + x_2$, $x \in R^2$; then $F(Y) = Y_1^2 + Y_2$, $Y \in I_\infty^2$, is nonwasteful. If $f(x) = \sin x$, $x \in R$, $F(Y) = [-1, 1]$, if $Y \in I$, and $F(Y) = [-2, 2]$ if $Y \in I_\infty \backslash I$ then F is wasteful.

Since there is no danger of misunderstanding we also write f and F instead of f_0 and F_0 in the sequel.

The *global unconstrained optimization problem over unbounded domains* can now be written concisely as follows. Let $X \in I_\infty^m$ and $f: X \to R$ be continuous. The problem to be solved is

$$\min_{x \in X} f(x). \tag{5}$$

Problem (5) is reduced to the *compactified problem*, which is

$$\min_{x \in \bar{X}} f(x). \tag{6}$$

This minimum always exists if the extension of f on $\bar{X}$ is defined via (1). The following algorithm aims to determine the global minimum f^+ and X^+, the set of global minimizers of problem (6).

Algorithm 2 will be syntactically equal to Algorithm 1, but now, unbounded compactified boxes $\bar{X}$ of $X \in I_\infty^m$, functions $f: \bar{X} \to \bar{R}$ and inclusion functions $F: \bar{I}(\bar{X}) \to \bar{I}$ are admitted as input data. We use the formulas for width and midpoint as they have just been introduced.

As Algorithm 1 does, Algorithm 2 produces, at the nth iteration, a list L_n consisting of pairs (Z_{ni}, z_{ni}), $i = 1, \cdots, l_n$, where l_n is the list length and $z_{ni} = \mathrm{lb}\, F(Z_{ni})$. The leading pair of L_n is denoted by (Y_n, y_n), and $(\tilde{Y}_n, \tilde{y}_n)$ denotes a pair of L_n satisfying $\tilde{y}_n \leqq z_{ni}$ for $i = 1, \cdots, l_n$. The function value $f_n \in \bar{R}$ is the lowest value of f produced up to the nth iteration. As before, $U_n = \bigcup_i Z_{ni}$, and (U_n) is a nested sequence.

4. Convergence properties of Algorithm 2. In this section, assumptions are looked for under which Algorithm 2 converges to the solution set of the compactified problem (6). First of all,

$$w(Y_n) \to 0 \quad \text{as } n \to \infty. \tag{7}$$

The proof is similar to the proof of the bounded case in [14] and is suppressed. Also, from the execution of Algorithm 2, we have

$$\tilde{y}_n \leqq f^+ \leqq f_n \quad \text{for all } n. \tag{8}$$

Let $a = (a_1, \cdots, a_m) \in \overline{R^m}$, and A, B be compact subsets of $\overline{R^m}$. We define

$$d_0(a, B) := \min_{b \in B} \left(\sum_{i=1}^{m} w([\min(a_i, b_i), \max(a_i, b_i)]^2 \right)^{1/2},$$

$$d_0(A, B) := \max_{a \in A} d_0(a, B),$$

$$d(A, B) := \max\{d_0(A, B), d_0(B, A)\}.$$

d is some kind of distance and is a means for describing convergence of sequences such as (U_n) to X^+ with respect to the natural topology of $\overline{R^m}$. Within $[-\lambda, \lambda]^m$, d coincides with the usual Hausdorff-metric for compact sets. We write $A_n \to B$ instead of $d(A_n, B) \to 0$ in the sequel. The notation $A_n \to B$ is consistent with the convergence of point sequences, if $A_n, B \in R^m$.

Again let $X \in I_\infty^m$, let $f: \bar{X} \to \bar{R}$ satisfy (1), and let $F: \bar{I}(\bar{X}) \to \bar{I}$ be a nonwasteful inclusion function for f satisfying (3). We consider the assumption

$$w(F(Y)) \to 0 \quad \text{as } w(Y) \to 0 \quad \text{for } Y \in I(X), \tag{9}$$

which implies continuity of f over X such that (6) has a solution, i.e., $X^+ \neq \emptyset$ and f^+ exists. Assumption (9) forced Algorithm 1 to converge if X was bounded [11]. We will see that (9) is also an appropriate convergence condition in the unbounded case. If Algorithm 2 (with or without monotonicity test) is applied to F, f, and $\bar{X}$, then the following theorem holds for the output data.

THEOREM 1. *If* (9) *holds, then as* $n \to \infty$,

(i) $f^+ - \tilde{y}_n \to 0$ *as well as* $f^+ - y_n \to 0$,

(ii) $f_n - f^+ \to 0$ *where* $\tilde{y}_n \leqq f^+ \leqq f_n$,

(iii) $U_n \to X^+$ *where* $X^+ \subseteq U_n$.

Proof. (i) We show that $f^+ - \tilde{y}_n \to 0$. The second assertion, $f^+ - y_n \to 0$ will follow from (ii), since $\tilde{y}_n \leqq y_n \leqq f_n$ holds.

If $f^+ = -\infty$ then $\tilde{y}_n = f^+$ because of (8). Let $f^+ \in R$. We focus on the nth iteration for one moment. Since F is nonwasteful, we have

$$F(\tilde{Y}_n) = \overline{\bigcup_{i \in J_n} F(B_{ni})}$$

for some partition $\tilde{Y} \cap R^m = \bigcup_{i \in J_n} B_{ni}$. Since $\tilde{y}_n = \text{lb } F(\tilde{Y}_n)$ and since $f^+ \in F(B_{ni})$ or $f^+ < \text{lb } F(B_{ni})$ for any $i \in J_n$, the following choice is possible.

If $\tilde{y}_n = -\infty$ we choose B_{ni_n}, $i_n \in J_n$, such that

$$\text{lb } F(B_{ni_n}) < f^+ - 2^n. \tag{10}$$

If $\tilde{y}_n \in R$ we choose B_{ni_n}, $i_n \in J_n$, such that either $\tilde{y}_n \in F(B_{ni_n})$ or both, $\text{lb } F(B_{ni_n}) - \tilde{y}_n < 2^{-n}$ as well as $\text{lb } F(B_{ni_n}) - f^+ < 2^{-n}$ holds.

Now, (7) implies $w(B_{ni_n}) \to 0$ such that $w(F(B_{ni_n})) \to 0$ follows. By (10), $\tilde{y}_n \in R$ holds for sufficiently large n, which means that $\tilde{y}_n \to f^+$.

(ii) We must show that $f_n \searrow f^+$. Let us, however, assume that $f_n \searrow \alpha$ for some $a > f^+$. Since $f^+ = f(x^+)$ for some $x^+ \in X^+$ and since f satisfies (1), there exists a sequence (ξ_k), $\xi_k \in X \subseteq R^m$, such that $\xi_k \to x^+$ and $f(\xi_k) \to f(x^+)$ as $k \to \infty$. Let k be fixed such that $f(x_k) < \alpha$. There exists a sequence (Z'_n) with $\xi_k \in Z'_n$ and Z'_n belonging to L_n, for any n. The existence of (Z'_n) is guaranteed since ξ_k is never excluded by the midpoint test. Condition (7) implies $w(Z'_n) \to 0$, and further $Z'_n \to \xi_k$. Thus, $Z'_n \in I(X)$, for sufficiently large n. We apply (9) and get $w(F(Z'_n)) \to 0$ and $F(Z'_n) \to f(\xi_k)$. For some large n we thus have $\text{ub } F(Z'_n) < \alpha$. Since $f_n \leqq \text{ub } F(Z'_n)$, we have a contradiction.

(iii) Since $X^+ \subseteq U_n$ is obvious, we show that $x \in U_n$, for all n, implies $x \in X^+$. We assume that $f(x) > f^+$ in order to get a contradiction. Since x occurs in every list, a sequence (Z'_n) exists where $x \in Z'_n$ and Z'_n belongs to L_n. Since F is nonwasteful, a partition of $Z'_n \cap R^m$ exists, $Z'_n \cap R^m = \bigcup_{i \in J_n} B_{ni}$. If $Z'_n \in I(X)$, set $B_{ni_n} := Z'_n$. Otherwise choose $i_n \in J_n$ such that $\text{lb } F(B_{ni_n}) \leqq \text{lb } F(Z'_n)$ or that $\text{lb } F(B_{ni_n})$ is asymptotically close to $\text{lb } F(Z'_n)$. Since $w(Z'_n) \to 0$, it follows that $B_{ni_n} \to x$. Let $\xi_n \in B_{ni_n}$; then $f(\xi_n) \to \alpha$ for some $\alpha \geqq f(x)$ due to (1). Now, $w(B_{ni_n}) \to 0$ implies $w(F(B_{ni_n})) \to 0$ and $\text{lb } F(B_{ni_n}) \to \alpha$, and further, $\text{lb } F(Z'_n) \to \alpha$. By (ii), we have $f_n < f(x) \leqq \alpha$ for large n such that Z'_n and thus x is discarded by the midpoint test. This gives the contradiction. □

Condition (9) is too restrictive for functions that have unbounded ranges. Let, for example, $f(x)=x^2$ and $F(Y)=\Box f(Y)$ be the inclusion function. Although the assertions (i)-(iii) of Theorem 1 are satisfied, condition (9) does not hold. In such cases the condition $w(F(Y))-w(\Box f(Y))\to 0$ as $w(Y)\to 0$, for $Y\in I(X)$ could be appropriate. In practice, however, this condition is too difficult to verify. Therefore we establish a condition with which we have obtained the best practical results. It combines theoretical as well as practical requirements where the computer may verify the latter for us.

Let $X\in I_\infty^m$, $f:\bar{X}\to R$, and an inclusion function $F:\bar{I}(\bar{X})\to\bar{I}$ of f be given. We assume that for any given $Z\in I(X)$

$$w(F(Y))\to 0 \quad \text{as } w(Y)\to 0 \quad \text{for } Y\in I(Z). \tag{11}$$

Condition (11) is not at all restrictive. If, for instance, f is continuous and programmable and if F is constructed via natural interval extensions (cf. § 6) then (11) is already satisfied. Then $X^+\neq\varnothing$ and f^+ exists. If Algorithm 2 is applied to F, f, and $\bar{X}$, the following theorem holds for the output data.

THEOREM 2. *If* (11) *holds and if there exists a number n such that the list* L_n *contains only bounded boxes, propositions* (i)-(iii) *of Theorem* 1 *are valid.*

Proof. Let Z be the smallest box of I^m that contains the boxes of L_n, that is, $X^+\subseteq U_n\subseteq Z$. We can now think of L_n as a list created by applying Algorithm 1 to f, F, and Z, such that the assertion of the theorem follows from the properties of Algorithm 1, cf. [11]. □

Example. Let us consider the well-known six-hump-camel-back function, $f(x)=4x_1^2-2.1x_1^4+x_1^6/3+x_1x_2-4x_2^2+4x_2^4$, $x\in X=R^2$. We compare two inclusion functions, F and F_1 of f over $\bar{X}$:

$$F(Y):=4Y_1^2+Y_1^4(Y_1^2/3-2.1)+Y_1Y_2+4Y_2^2(Y_2^2-1)$$

$$F_1(Y):=4Y_1^2-2.1Y_1^4+Y_1^6/3+Y_1Y_2-4Y_2^2+4Y_2^4 \qquad \text{for } Y\in I_\infty(X).$$

$$F(\bar{Y}):=\overline{F(Y)}, \qquad F_1(\bar{Y}):=\overline{F_1(Y)}$$

Neither F nor F_1 satisfy (9). Both F as well as F_1 satisfy (11). If Algorithm 2 is applied to f (using (1)), $\bar{X}$ and F then, after a few iterations, the lists L_n do not contain any unbounded boxes. The assumptions of Theorem 2 are therefore computationally verified for F (cf. Example 1 of § 8 for computational results). This is not the case if F_1 is chosen, since each list will contain unbounded boxes (as long as the monotonicity test is not used). Let, for instance, $\bar{Y}=[a,\infty]\times[b,\infty]$ for $a,b\in R$; then $F_1(\bar{Y})=\bar{R}$ (cf. § 6). Hence $\bar{Y}$ is never discarded by the midpoint test. Nevertheless, the sequence (U_n) will converge to a superset of X^+ that contains the points $(\pm\infty,\pm\infty)$, and we have $\tilde{y}_n=-\infty$ for any n.

Remarks. (1) It is difficult to present precise conditions for F to satisfy the bounded box assumption of Theorem 2. In practice, this assumption was satisfied if F was constructed so that $-\infty\notin F(\bar{Y})$ for all boxes $Y=Y_1\times\cdots\times Y_m\in I_\infty(X)$ where $Y_i=(-\infty,-a]$ or $Y_i=[a,\infty)$, for some arbitrarily large real $a>0$.

(2) The termination criteria known for Algorithm 1 [3], [11] are also appropriate for Algorithm 2 by extending the criteria to infinite boxes and numbers.

Let us return to problem (5) originally posed. Let X^* be the set of global minimizers of problem (5) and f^* the global minimum. Algorithm 2 is appropriate to solve (5) via (6) using the following theorem whose proof is obvious.

THEOREM 3. (i) *If* $f^+=-\infty$, *then* (5) *has no solution, and f is unbounded from below in X.*

(ii) *If* $f^+ \in R$ *and if* $X^* := X^+ \cap R^m$ *is nonempty, then* X^* *and* $f^* = f^+$ *is the solution of* (5). *If* $X^* = \emptyset$ *then* (5) *has no solution but* f *is bounded from below.* □

5. The monotonicity test. The well-known monotonicity test (3), (10) is extended to the unbounded case. This is even more important in the unbounded case, as functions are frequently strictly monotone for large values of the variables, such as polynomials. Such areas contain no minimizers and can be discarded from the lists.

Let $X \in I_\infty^m$, $f: X \to R$, and let $\partial f_i(x)$ be the ith component of the generalized gradient of f at x that is defined as

$$\partial f(x) = \operatorname{conv}\left\{\lim_{n\to\infty} f'(x_n)\colon x_n \to x,\ x_n \in X,\ x_n \notin S \cup \Omega\right\} \tag{12}$$

if at least one limit exists. Here, conv denotes the convex hull, $f'(x_n)$ the gradient of f at x_n, Ω the set of points in some neighbourhood of x at which f is not differentiable, and S any set of Lebesgue measure 0 (cf. Clarke [2]).

Let $Y \in \bar{I}(\bar{X})$, and let $\partial f(x)$ exist for every $x \in Y \cap R^m$ and $G_i(Y) \in \bar{I}$ an inclusion of ∂f_i in Y in the sense that $\partial f_i(x) \in G_i(Y)$ for any $x \in Y \cap R^m$. (In this connection, X cannot be replaced by Y in (12).) We set $Y_i = [a_i, b_i] \in \bar{I}$ and $\bar{X}_i = [c_i, d_i] \in \bar{I}$. Furthermore, let $Y(i/s)$ for $s \in Y_i$ denote that box that arises from Y by replacing Y_i with s. Then the *monotonicity test* (destined to handle the boxes $Y = V_1$, V_2 of step (10) of Algorithm 2) consists of the following two parts:

(I) Test for strictly monotone increasing. For some $i = i, \cdots, m$, if $0 < \operatorname{lb} G_i(Y)$ then

(i) If $c_i < a_i$ then discard (Y, y) from the list.

(ii) If $c_i = a_i \in R$ then replace (Y, y) with the pair (Y', y') where $Y' = Y(i/a_i)$ and $y' = \operatorname{lb} F(Y')$.

(iii) If $a_i = -\infty$ then terminate Algorithm 2 (since $f^+ = -\infty$ such that problem (5) has no solution).

(II) Test for strictly monotone decreasing. Analogous to (I).

Remarks. (1) It is a consequence of a termination by (iii) that only one global minimizer of X^+ is found. But, this is sufficient to guarantee the unsolvability of (5) (cf. Theorem 3).

(2) It is favourable to admit $\pm\infty$ as values of the limits in the definition of the generalized gradient (12). This is shown in the following example where even the bounded case is improved.

Example. Let the semicircle $f(x) = (1-x^2)^{1/2}$ be defined on $X = [-1, 1]$. Then, $X^* = \{-1, 1\}$ and $f^* = 0$. Note that $f'(x) \to \mp\infty$ as $x \to \pm 1$. We applied Algorithm 2 with monotonicity test to this problem, considering Remark (2). We took $F(Y) = (1 - Y^2)^{1/2}$ and $G(Y) = -Y(1 - Y^2)^{1/2}$ as inclusion functions for f and ∂f, respectively, and got the exact result after three iterations. See the next section for the computation of $G(Y)$.

6. How to get inclusion functions. The *simplest* way to get the inclusion functions that are necessary for Algorithm 2 to run is, first, to develop a calculus in I_∞ and to construct inclusion functions with respect to I_∞^m via Moore's principle of natural interval extensions [9], and second, to extend these inclusion functions to $\bar{I}^m$ by means of (3) and (4). We do not use an arithmetic in $\bar{I}$ (cf. [5]), to get the inclusions required since the resulting intervals would be too large in this case.

Let $A, B \in I_\infty$. We expect the arithmetic in I_∞ to satisfy

$$A * B = \{a * b\colon a \in A, b \in B\} \tag{13}$$

if $*$ stands for $+$, $-$, and $\cdot$ (product), and A/B is the smallest interval of I_∞ or is the union of the two smallest intervals of I_∞ such that

$$A/B \supseteq \{a/b\colon a \in A,\ b \in B,\ b \neq 0\}. \tag{14}$$

The case $A/0$ is excluded.

For example, $1/[1, \infty) = [0, 1]$, or $1/[-1, 1] = (-\infty, -1] \cup [1, \infty)$. Hence, I_∞ is not closed with respect to division. We do not worry about that, and we split up the union that occurs into two intervals of I_∞ and process the two intervals separately as long as necessary. For example,

$$\begin{aligned} 1/[-1, 1] + [0, 2]/[1, 2] &= ((-\infty, -1) + [0, 2]) \cup ([1, \infty) + [0, 2]) \\ &= (-\infty, 1] + [1, \infty) = \boldsymbol{R} \in I_\infty. \end{aligned}$$

It is not difficult to establish explicit formulas for the arithmetic defined by (13) and (14) such that we abstain from giving further details. The best way to derive these formulas is to apply limit operations to Moore's formulas for the bounded arithmetic [9], [10]. For instance,

$$[0, 1](-\infty, 0] = \lim_{a \to -\infty} [0, 1][a, 0] = (-\infty, 0].$$

If $f\colon D \to R$, $D \subseteq R^k$ is a function predeclared in the programming language used (such as sin, cos, etc.) and if $Y \in I_\infty^k$, then the *natural interval extension* of f to Y denoted by $f(Y)$ is defined as the smallest interval of I_∞ such that

$$f(Y) \supseteq \square f(Y \cap D). \tag{15}$$

Practically, this definition causes no trouble since the ranges of the functions usually predeclared are well known. For example, if $Y = [-1, \infty)$, then $\sin Y = [-1, 1]$, $\ln Y = R$, and $\exp(-Y) = [0, e]$.

It is necessary to admit boxes Y in (15) that are not necessarily contained in the domain of f since at each step of a recursive evaluation of a natural interval extension, an overestimation of the range is likely. Hence the domain of the function of the next recursive step can be exceeded. Such a superfluous overestimate is prevented by the intersection $Y \cap D$.

Finally, a *natural interval extension* of any programmable function f over $Y \in I_\infty(X)$, $X \in I_\infty^m$ can be defined recursively via (13)-(15), in the same way a function value $f(x)$ is defined recursively via the basic functions (arithmetic operations, predeclared functions), for example, by means of a computer code. The recursive representation of programmable functions is treated in detail in [8], [13], and elsewhere.

Example. If $f(x) = \exp(1/(|\sin x| + |\cos x|))$ and if $Y = R$, then the natural interval extension of f to R is $f(R) = \exp(1/(|\sin R| + |\cos R|)) = [\sqrt{e}, \infty)$.

The result gained in this example depends on the representation for $f(x)$ chosen (cf. also the example after Theorem 2). Thus we can see that it is important to choose appropriate function representations. A procedure to construct them and further details can be found in [14].

7. The realization on the computer. There exist several programming languages and software packages that are able to realize Moore's interval arithmetic on a computer. They also control the rounding errors such that logical flaws cannot occur [3], [6], [9]. Up to now there has not been a general widespread programming language in which an infinite interval arithmetic is incorporated. This is, however, no real problem for a programmer. In our case, we must be aware that both kinds of infinite intervals that we deal with must be simulated by intervals on the computer.

A simple procedure we used is the following. Let R_M be the set of machine representable real numbers. We assume that $\pm L$ is the largest (smallest) number of R_M. Let $I_M = \{[a, b]: a, b \in R_M, a \leqq b\}$ be the set of *machine intervals*. We call $[a, b] \in I_M$ *machine-finite* if $|a|, |b| < L$, otherwise *machine-infinite*. Now, the intervals that occur at the execution of Algorithm 2 on a machine must simulate the intervals that occur at the exact (nonmachine) execution. Hence, a machine-finite interval $[a, b]$ simulates and means $[a, b]$ itself. A machine-infinite interval such as $[A, L]$ means either $[a, \infty)$ if inclusion functions are constructed or $[a, \infty]$ if the execution of Algorithm 2 is addressed. The situation seems involved but is not. It even has the great practical advantage that, when transmitting the inclusion $F(Y)$ to $F(\bar{Y})$, the simulating machine intervals need not be changed.

Width and midpoint of machine intervals are defined as width and midpoint of the corresponding simulated intervals. For instance, $w([1, L]) = \lambda^2$.

An arithmetic $\tilde{*}$ in I_M where $*$ stands for $+$, $-$, $\cdot$, and $/$ is defined as follows. Let $A, B \in I_M$; then $A \tilde{*} B$ is the smallest interval of I_M or the union of the smallest two intervals of I_M such that $A \tilde{*} B \supseteq A * B$ (more precisely, such that $A \tilde{*} B$ is the smallest interval of I_M that *simulates* an inclusion of $A * B$, etc.). In case of division, $B = 0$ is excluded. For instance,

$$[2, L] \tilde{+} [-L, L] = [-L, L] \quad \text{or} \quad [2, L] \tilde{:} [-L, -2] = [-L, -4].$$

Such arithmetics can be fulfilled easily (cf. [6]).

The approximation of the interval values for the functions predeclared by intervals of I_M is done straight, for instance, $\ln [-1, L] = [-L, L]$.

If now Algorithm 2 runs on a machine, then after the computation is terminated either the information $f^+ = -\infty$ (due to the monotonicity test) will be delivered or a machine interval $A \supseteq [\tilde{y}_n, f_n]$ and machine boxes $W_i \supseteq Z_{ni}$, $i = 1, \cdots, l_n$, will be the output data. Here n is the final iteration index and l_n is the length of the list L_n. In general, we get inclusion A and W_i instead of $[\tilde{y}_n, f_n]$ and Z_{ni} because of the common outward rounding when a machine interval arithmetic is used. Due to Theorem 3, the output data of the machine-computation with Algorithm 2 must be interpreted as follows to obtain the required solution of problem (5):

(1) If $f^+ = -\infty$ then f is unbounded from below and (5) has no solution.

(2) If A and W_i, $i = 1, \cdots, l_n$, are machine-finite, then problem (5) has a solution, f^* and X^*, and $f^* \in A$, $X^* \subseteq W := \bigcup_{i=1}^{l_n} W_i$.

(3) If A or if W_{ni} for at least one $i \in \{1, \cdots, l_n\}$ is machine-infinite, then a decision has not been possible whether or not a solution of (5) does exist. However, if a solution exists, f^* and X^*, then $f^* \in A \cap R$ and $X^* \subseteq W \cap R^m$.

8. Numerical results. The following examples were executed in PASCAL-SC.

Example 1. Six-hump-camel-back function, $f(x) = 4x_1^2 - 2.1x_1^4 + x_1^6/3 + x_1 x_2 - 4x_2^2 + 4x_2^4$ for $x \in R^2$. We used the inclusion function F as described in the example after Theorem 2 for larger boxes Y and the mean-value form (cf. [15]) for F if the boxes Y were machine-finite with $w(Y) \leqq 1$. Starting box was $\bar{R}^2$, respectively, $[-L, L]^2$. We needed 211 iterations of Algorithm 2 with monotonicity test (which is about 422 interval function evaluations of $F(Y)$) to obtain the intended absolute accuracy of 10^{-6} for the solution, $f^* \in -1.03162\,84535\,8 + [0, 5]10^{-11}$, $X^* \subseteq W_1 \cup W_2$, where

$$W_1 = [-8.98426\,8, -8.98414\,8]10^{-2} \times [7.12655\,78, 7.12656\,98]10^{-1},$$

$$W_2 = -W_1.$$

By contrast, when we applied Algorithm 1 (also with monotonicity test) with starting box $[-2.5, 2.5]^2$ we needed 163 iterations (about 326 evaluations of $F(Y)$).

Example 2. Wolfe's [18] function modified by Zowe [19] is defined as follows:

$$f(x)=\begin{cases}5(9x_1^2+16x_2^2)^{1/2} & \text{if } x_1 \geqq |x_2|,\\ 9x_1+16|x_2| & \text{if } 0<x_1<|x_2|,\\ 9x_1+16|x_2|-x_1^9 & \text{if } x_1 \leqq 0,\end{cases}$$

where $x \in R^2$. The only global minimizer of f with respect to R^2 is $x^*=(-1,0)$ and $f^*=-8$. The function is convex, and f fails to be differentiable only on the ray $x_1 \leqq 0$, $x_2=0$. As inclusion functions we used natural interval extensions and their unions (if Y was assigned to more than one function branch) for larger and unbounded boxes, otherwise we used the mean value form of f on Y (cf. [14]) where inclusions of the generalized gradient, instead of inclusion of the derivative, were taken if no derivative was available.

When we applied Algorithm 2 (with monotonicity test) to f, F, and $\bar{X}=\bar{R}^2$, 106 iterations (about 212 evaluations of $F(Y)$) were needed to determine the solutions x^* and f^* within an absolute accuracy of $2 \cdot 10^{-6}$.

Example 3. Let $X=[-1,1]\times[-1,1]\times[0,\infty) \subseteq R^3$ and $f: X \to R$ be defined by

$$f(x)=(1-x_1^2)^{1/2}\cos x_3+(1-x_2^2)^{1/2}/(1+x_3^2)+2x_3\, e^{-x_1}.$$

There exist four global minimizers of f in X having the coordinates $x_1=\pm 1$, $x_2=\pm 1$, $x_3=0$. The objective function f is differentiable in the interior of X, continuous—but not Lipschitz—on X. However, f is generalized differentiable (when infinite values are admitted) on the edge of X that contains the four minimizers. As inclusion functions we used the plain natural interval extension as well as the mean-value form. Algorithm 2 (with monotonicity test) needed 31 iterations (which makes about 62 evaluations of $F(Y)$) to achieve the intended absolute accuracy of about 10^{-6} for X^* and of 10^{-10} for f^*.

Example 4. One of the unknown referees suggested considering the function $f(x)=(x_1-x_2)^2$. The set of global minimizers X^* consists of the whole line $x_1=x_2$. When we applied Algorithm 2 (with monotonicity test) to f, $F(Y)=(Y_1-Y_2)^2$, $\bar{X}=\bar{R}^2$, and $\lambda=10$, we needed 733 iterations to obtain an intended maximum box width of one (which was chosen to avoid a too long computer run). The final list contained 313 bounded and five unbounded boxes that were covering the line $x_1=x_2$. The final inclusion for the minimum value $f^*=f^+$ was $[0,0]$.

Acknowledgments. The authors are greatly indebted to Tom Heilandt and R. Ernemann for programming and computing the examples in § 8.

REFERENCES

[1] N. S. ASAITHAMBI, Z. SHEN, AND R. E. MOORE, *On computing the range of values*, Computing, 28 (1982), pp. 225-237.

[2] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[3] E. R. HANSEN, *Global optimization using interval analysis—the multidimensional case*, Numer. Math., 34 (1980), pp. 247-270.

[4] K. ICHIDA AND Y. FUJII, *An interval arithmetic method for global optimization*, Computing, 23 (1979), pp. 85-97.

[5] W. M. KAHAN, *A more complete interval arithmetic*, Lecture notes for a summer course at the University of Michigan, Ann Arbor, MI, 1968.

[6] U. KULISCH AND W. L. MIRANKER, *Computer Arithmetic in Theory and Practice*, Academic Press, New York, 1981.

[7] S. E. LAVEUVE, *Definition einer Kahan-Arithmetik und ihre Implementierung*, in Interval Mathematics, K. Nickel, ed., Springer-Verlag, Heidelberg, 1975, pp. 236-245.

[8] G. P. MCCORMICK, *Nonlinear Programming: Theory, Algorithms and Applications*, John Wiley, New York, 1983.
[9] R. E. MOORE, *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1966.
[10] ———, *Methods and Applications of Interval Analysis*, Studies in Applied Math 2, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.
[11] R. E. MOORE AND H. RATSCHEK, *Inclusion functions and global optimization* II, Math. Programming, 41 (1988), pp. 341-356.
[12] P. Y. PAPALAMBROS AND D. WILDE, *Principles of Optimum Design*, University of Cambridge Press, Cambridge, 1988.
[13] L. B. RALL, *Automatic Differentiation: Techniques and Applications*, Springer-Verlag, New York, 1981.
[14] H. RATSCHEK, *Inclusion functions and global optimization*, Math. Programming, 33 (1985), pp. 300-317.
[15] H. RATSCHEK AND J. ROKNE, *Computer Methods for the Range of Functions*, Horwood, Chichester, 1984.
[16] ———, *Efficiency of a global optimization algorithm*, SIAM J. Numer. Anal., 24 (1987), pp. 1191-1201.
[17] S. SKELBOE, *Computation of rational interval functions*, BIT, 14 (1974), pp. 87-95.
[18] P. WOLFE, *A method of conjugate subgradients for minimizing nondifferentiable functions*, Math. Programming Stud., 3 (1975), pp. 145-173.
[19] J. ZOWE, *Nondifferentiable optimization*, in Computational Mathematical Programming, K. Schittkowski, ed., Springer-Verlag, Berlin, 1985, pp. 321-356.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 540-551, May 1990

004

GLOBAL CONTROLLABILITY OF LOCALLY LINEARIZABLE SYSTEMS*

R. M. HIRSCHORN†

Abstract. A nonlinear control system that is locally feedback linearizable usually will not have the global controllability properties of a linear system. The purpose of this paper is to study the reachable set for such systems and to generate controls to accomplish desired state transfers.

Key words. controllability, affine systems

AMS(MOS) subject classification. 93B05

1. Introduction. There has been considerable interest in classifying those nonlinear systems that are locally feedback linearizable (cf. Brockett [1], Jakubczyk and Respondek [2], and [3]-[6]). These ideas have recently been employed to control the motion of robot manipulators (cf. [7]-[9]). The purpose of this paper is to study the global controllability of such systems.

A nonlinear control system is locally feedback linearizable if, replacing u by $(\beta(x)u+k(x))$ and changing coordinates in some open neighbourhood of the initial state, we can obtain a controllable time-invariant linear system. For the nonlinear system $\dot{x}=f(x)+ug(x)$; $x(0)=x_0$ with $f(x_0)=0$, necessary and sufficient conditions for feedback linearization have been derived based on the way f and g generate a Lie algebra of vector fields (cf. [1]-[3]). It is natural to suppose that systems which satisfy these Lie algebraic criteria exhibit some of the standard controllability properties associated with linear control systems, for example, the reachable set of states includes an open neighbourhood of the initial state. The following example shows that this is not always the case if $f(x_0)\neq 0$.

Example 1.1. Consider the system model

$$\dot{x}_1(t)=u(t), \qquad \dot{x}_2(t)=e^{x_1(t)} \tag{1.1}$$

with $\underline{x}(0)=\underline{x}_0=(0,0)$ and $\underline{x}\in M=R^2$. This system is affine with $f(x_1,x_2)=(0,e^{x_1})$, and $g(x_1,x_2)=(1,0)$ and $\{f,g\}$ satisfy the Lie algebraic criteria for local feedback linearization, although $f(x_0)\neq(0,0)$. Note that rescaling u by $\tilde{u}=e^{x_1}u$ followed by the change in coordinates $z_1=e^{x_1}$, $z_2=x_2$ we obtain the *controllable* linear system

$$\begin{bmatrix}\dot{z}_1\\ z_2\end{bmatrix}=\begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix}\begin{bmatrix}z_1\\ z_2\end{bmatrix}+\begin{bmatrix}1\\ 0\end{bmatrix}\tilde{u}, \qquad \underline{z}(0)=\underline{z}_0=(1,0). \tag{1.2}$$

Unfortunately the original system (1.1) is not even locally controllable about $\underline{x}_0$ since at any time $t>0$ the reachable set of states is

$$\mathcal{R}_t(\underline{x}_0)=R\times(0,\infty)$$

and thus $\underline{x}_0$ is on the boundary of the reachable set.

This example illustrates that, for systems with the Lie algebraic properties for feedback linearization operating from a point where $f(x_0)\neq 0$, controllability depends on more than the structure of the Lie algebra of vector fields generated by f and g. The geometry of the state manifold and the integral curves of f and g also play a role.

* Received by the editors November 4, 1987; accepted for publication (in revised form) May 30, 1989.
† Department of Mathematics and Statistics, Queen's University, Kingston, Ontario K7L 3N6, Canada.

In § 2 a globally defined output map h is associated with systems that satisfy the Lie algebraic criteria for local feedback linearization. A set $\mathcal{Y}_h(x_0)$ of admissible C^∞ output functions is described. This set includes all outputs that result from C^∞ inputs. In § 3 the set of reachable states is deduced from $\mathcal{Y}_h(x_0)$, and output tracking is used to explicitly define a control u to transfer the state of the system from $x(t_0)=x_0$ to $x(t_1)=x_1$. A class of systems that are globally controllable is identified.

2. State transfers using output tracking. For linear control systems the variation of constants formula can be utilized to explicitly find a control function u that will transfer the state from $x(t_0)=x_0$ to $x(t_1)=x_1$ $(t_1>t_0)$ (cf. Brockett [10]). Another approach to generating u is to use output tracking. For example, consider the linear system

$$\dot{x}(t)=Ax(t)+bu(t), \qquad x\in R^n. \tag{2.1}$$

Let $\alpha=\operatorname{rank}[b, Ab, \cdots, A^{n-1}b]$ and choose c to be any $1\times n$ matrix over R with

$$cb=cAb=\cdots=cA^{\alpha-2}b=0$$

and $cA^{\alpha-1}b\neq 0$.

If $y(t)=cx(t)$ is considered to be an output for (2.1), then

$$\begin{aligned}
&y^{(1)}(t)=\dot{y}(t)=c\dot{x}(t)=cAx(t),\\
&y^{(2)}(t)=cA^2x(t),\\
&\vdots\\
&y^{(\alpha-1)}(t)=cA^{\alpha-1}x(t),\\
&y^{(\alpha)}(t)=cA^{\alpha}x(t)+cA^{\alpha-1}bu(t).
\end{aligned} \tag{2.2}$$

Let

$$c^\alpha(x)=\begin{bmatrix} cx\\ cAx\\ \vdots\\ cA^{\alpha-1}x\end{bmatrix} \quad\text{and}\quad y^\alpha(t)=\begin{bmatrix} y(t)\\ y^{(1)}(t)\\ \vdots\\ y^{(\alpha-1)}(t)\end{bmatrix}$$

so that when $x(t)$ is a trajectory for (2.1) then $y^\alpha(t)=c^\alpha(x(t))$. It is easy to check that c^α is a 1-1 map of Range $[b, Ab, \cdots, A^{\alpha-1}b]$ onto R^α. Thus, to transfer the state of (2.1) from $x(t_0)=0$ to $x(t_1)=x_1\in$ Range $[b, Ab, \cdots, A^{\alpha-1}b]$ is equivalent to transferring the output $y(\cdot)$ from $y^\alpha(t_0)=c^\alpha(x_0)=c^\alpha(0)=\underline{0}$ to $y^\alpha(t_1)=c^\alpha(x(t_1))=c^\alpha(x_1)$.

Now choose any function $y_d(t)$ with $y_d^\alpha(t_0)=0$ and $y_d^\alpha(t_1)=c^\alpha(x_1)$. From (2.2)

$$y_d^{(\alpha)}(t)=cA^\alpha x(t)+cA^{\alpha-1}bu(t),$$

and thus using the control

$$u_d(x,t)=\frac{y_d^{(\alpha)}(t)-cA^\alpha x}{cA^{\alpha-1}b},$$

the output $y(t)$ has $y^\alpha(t_0)=0$, $y^\alpha(t_1)=c^\alpha(x_1)$. In particular, using u_d it follows that $x(t_0)=0$ and $x(t_1)=x_1$. This method of achieving state transfers will be generalized to nonlinear systems that, in essence, are locally feedback linearizable.

Consider the single-input nonlinear system model

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)); \qquad x(t_0) = x_0 \in M \tag{2.3}$$

where M is a C^∞ (smooth) n-dimensional manifold, f, g are C^∞ vector fields on M and $u:[t_0, \infty) \to R$ is continuous.

Suppose $f(x_0) = 0$. Then the system (2.3) will be called *locally state linearizable on an open neighbourhood $\mathcal{U}_0$ of x_0* if the system can be transformed into a linear controllable system via a change of coordinates in $\mathcal{U}_0$ (cf. [6]). Suppose that (2.3) is modified on U_0 by replacing $u(t)$ by $(\beta(x(t))u(t) + k(x(t)))$ where $k(x)$ is a smooth feedback law defined on $\mathcal{U}_0$ with $k(x_0) = 0$ and $\beta(x)$ is a smooth change of coordinates in the input space with $\beta(x) \neq 0$ for all $x \in \mathcal{U}_0$. If the resulting system is locally state linearizable on $\mathcal{U}_0$ we say that (2.3) is *feedback linearizable on $\mathcal{U}_0$*.

If $f(x_0) = 0$ then a necessary and sufficient condition for the system (2.3) to be feedback linearizable on an open neighborhood $\mathcal{U}_0$ of x_0 in M is that the set of vectors

$$\{g(x), \operatorname{ad}_f g(x), \operatorname{ad}_f^2 g(x), \cdots, \operatorname{ad}_f^{n-1} g(x)\} \tag{2.4}$$

are linearly independent at each x in $\mathcal{U}_0$ and the distribution

$$\mathcal{D}(x) = \operatorname{span}\{g(x), \cdots, \operatorname{ad}_f^{n-2} g(x)\}$$

is involutive (cf. [1]-[3]). To study the structure of the reachable set for systems that satisfy (2.4) some global condition must be added. Note that when $f(x_0) \neq 0$ the system is not locally feedback linearizable. In this paper the existence of an output map $y = h(x)$ with relative order $\alpha_h = n$ (cf. [11]-[13]) will be assumed.

If $h \in C^\infty(M)$ is an output map $y = h(x)$ for (2.3), then the *relative order* α_h is the least nonnegative integer k such that $gf^{k-1}h \not\equiv 0$ on M, and $\alpha_h = \infty$ if $gf^k h \equiv 0$ for all $k \geqq 0$ (if X is a vector field and h is a function, then $Xh = dhX$). Thus for the linear system (2.1) with the output $y = h(x) = cx$ the relative order $\alpha_h = \alpha = \operatorname{rank}[b, Ab, \cdots, A^{n-1}b]$. If this system is controllable then $\alpha_h = n$. That is, for a controllable time-invariant linear system there are linear outputs with relative order $\alpha_h = n$. The following lemma shows that for nonlinear systems that satisfy condition (2.4) there are locally defined outputs with $\alpha_h = n$.

LEMMA 2.1. *Suppose that the nonlinear system* (2.3) *satisfies the Lie algebraic criteria* (2.4). *Then there exists an open neighborhood $\mathcal{U}_0$ of x_0 in M and a C^∞ output $y = h(x)$ defined on $\mathcal{U}_0$ with $\alpha_h = n (= \dim M)$.*

Proof. From (2.4) the distribution $\mathcal{D}(x) = \operatorname{span}\{g(x), \cdots, \operatorname{ad}_f^{n-2} g(x)\}$ is involutive on some open neighborhood $\mathcal{U}_{x_0}$ of x_0 in M, and $\operatorname{ad}_f^{n-1} g(x) \notin \mathcal{D}(x)$ for all $x \in \mathcal{U}_{x_0}$. The Frobenius Theorem (cf. [14]) asserts that there exist an open set $\mathcal{U}_0$ such that $x_0 \in \mathcal{U}_0 \subset \mathcal{U}_{x_0}$ and a coordinate system $(\mathcal{U}_0, x_1, \cdots, x_n)$ such that the slices $x_n = c$, c constant, are the integral manifolds of $\mathcal{D}$ around x_0. Define $h \in C^\infty(\mathcal{U}_0)$ to the function $h(x) = x_n$. Thus for all vector fields $X \in \mathcal{D}$, $Xh \equiv 0$. In particular, $gh = 0$,

$$\operatorname{ad}_f gh = (fg - gf)h = f(gh) - gfh = -gfh \equiv 0, \cdots, \operatorname{ad}_f^{n-2} gh$$

$$= (-1)^{n-3} gf^{n-2} h \equiv 0$$

and since $\operatorname{ad}_f^{n-1} g \notin \mathcal{D}$, $\operatorname{ad}_f^{n-1} gh = (-1)^{n-2} gf^{n-1} h \not\equiv 0$. This means that the relative order of h is $\alpha_h = n$, as required.

There can be obstructions to obtaining a global version of Lemma 2.1. The following (somewhat pathological) example illustrates this point.

Example 2.2. Consider the system (2.3) where $M = T^2$, the two-torus, and $f(x_1, x_2) = (1+2\cos 2\pi x_1, 0)$, $g(x_1, x_2) = (1, \sqrt{2})$, $x_0 = (0.5, 0.5)$, T^2 is modeled as the square $[0,1]\times[0,1]$ with the usual identifications. Since $\mathrm{ad}_f g(x_1, x_2) = (4\pi \sin 2\pi x_1, 0)$, it follows that (2.4) is satisfied and Lemma 2.1 applies. Thus locally there exist outputs h with $\alpha_h = 2$—for example, $\mathcal{U}_0 = (0,1)\times(0,1)$ and $h(x_1, x_2) = \sqrt{2}x_1 - x_2$ will suffice—but there is no $h \in C^\infty(T^2)$ with $\alpha_h = 2$! Suppose such an h exists. Then $gh \equiv 0$ implies that h is constant on the integral curves of g. Since each g-integral curve is a dense submanifold of T^2 (a skew line), h must be constant on T^2 by the continuity of h. In that case $gf^i h \equiv 0$ for all i so that $\alpha_h = \infty$, a contradiction.

In practice it is not unreasonable to expect the existence of a globally defined output function $y = h(x)$ with $\alpha_h = n$ when (2.4) holds.

As in the linear case the existence of outputs $y = h(x)$ with $\alpha_h = n$ gives a useful type of observability.

DEFINITION. Let $y = h(x)$ be a C^∞ output map for the system (2.3) with relative order α_h. Then define

$$y^\alpha(t) = \begin{bmatrix} y(t) \\ y^{(1)}(t) \\ \vdots \\ y^{(\alpha_h - 1)}(t) \end{bmatrix}$$

and

$$h^\alpha(x) = \begin{bmatrix} h(x) \\ fh(x) \\ f^2h(x) \\ \vdots \\ f^{\alpha_h - 1}h(x) \end{bmatrix}.$$

Since $y(t) = h(x(t))$,

$$\begin{aligned} y^{(1)}(t) &= dh_{x(t)}\dot{x}(t) \\ &= dh_{x(t)}(f(x(t)) + u(t)g(x(t))) \\ &= fh(x(t)) + u(t)gh(x(t)) \\ &= fh(x(t)) \\ &\vdots \\ y^{(\alpha_h - 1)}(t) &= f^{\alpha_h - 1}h(x(t)) \\ y^{(\alpha_h)}(t) &= f^{\alpha_h}(x(t)) + u(t)gf^{\alpha_h - 1}h(x(t)) \end{aligned}$$

we see that $y^\alpha(t) = h^\alpha(x(t))$. Thus if we observe $y^\alpha(t_1)$ we know $h^\alpha(x(t_1))$ which, in turn, implies some knowledge of the state $x(t_1)$.

LEMMA 2.2. *Suppose the system* (2.3) *satisfies the linearization condition* (2.4). *Then there exists an open neighborhood* $\mathcal{U}$ *of* x_0 *and a* C^∞ *output* $y = h(x)$ *defined on* $\mathcal{U}$ *with* $\alpha_h = n$, *and the map* $x \mapsto h^\alpha(x)$ *is a diffeomorphism of* $\mathcal{U}$ *into* R^n.

Proof (compare with [2], [3]). Lemma 2.1 implies the existence of an output h with $\alpha_h = n$ on $\mathcal{U}$. To complete the proof we need only show that $dh^\alpha_{x_0}$ is a linear

isomorphism—the Inverse Function Theorem then applies. Suppose $dh^{\alpha}_{x_0}v = \underline{0}$ for $v \in T_{x_0}(M)$. From (2.4), $v = a_0 g(x_0) + a_1 \operatorname{ad}_f g(x_0) + \cdots + a_{n-1} \operatorname{ad}_f^{n-1} g(x_0)$, and thus $dh^{\alpha}_{x_0}v = \underline{0}$ means that $dh_{x_0}v = 0$, $d(fh)_{x_0}v = 0, \cdots, d(f^{n-1}h)_{x_0}v = 0$. Using the first of these equalities, we have that $dh_{x_0}v = a_0 gh(x_0) + a_1 \operatorname{ad}_f gh(x_0) + \cdots + a_{n-1} \operatorname{ad}_f^{n-1} gh(x_0) = 0$ as $\alpha_h = n$, $gh \equiv \operatorname{ad}_f gh \equiv \cdots \equiv \operatorname{ad}_f^{n-2} gh \equiv 0$. From the proof of Lemma 2.1 we have $\operatorname{ad}_f^{n-1} gh(x_0) \neq 0$, so that $dh^{\alpha}_{x_0}v = 0$ implies $a_{n-1} = 0$. Similarly, $d(fh)_{x_0}v = 0$ implies $a_{n-2} = 0$, and continuing we can conclude that $v = 0$. Thus $dh^{\alpha}_{x_0}$ is a linear isomorphism and the proof is complete.

Remark. The output $y = h(x)$ described in Lemma 2.2 results in a realization that is *locally observable.* That is, if $x(t_1) \in \mathcal{U}$ and we observe $y^{\alpha}(t_1)(= h^{\alpha}(x(t_1)))$, then we can recover the state $x(t_1)$ via $x(t_1) = (h^{\alpha}|_{\mathcal{U}})^{-1}(y^{\alpha}(t_1))$. Knowing y and its derivatives at a fixed time enables us to deduce the state when $x \in \mathcal{U}$.

To study the structure of the reachable set for the system (2.3) a global version of Lemma 2.2 will be required.

DEFINITION. The system (2.3) can be *globally observed using the output* $y = h(x)$ if there exists a C^{∞} output $h(x)$ defined on M with $\alpha_h = n$, and the map $x \mapsto h^{\alpha}(x)$ is a diffeomorphism of M *into* R^n.

Suppose the system (2.3) can be globally observed using the output $y = h(x)$. To transfer the state from $x(t_0) = x_0$ to $x(t_1) = x_1$ $(t_1 > t_0)$ we can employ output tracking. If $x(t)$ is a trajectory for (2.3) then $y^{\alpha}(t) = h^{\alpha}(x(t))$. Thus to achieve the desired state transfer the output function $y(t)$ must satisfy

$$y^{\alpha}(t_0) = h^{\alpha}(x(t_0)) = h^{\alpha}(x_0), \qquad y^{\alpha}(t_1) = h^{\alpha}(x(t_1)) = h^{\alpha}(x_1). \tag{2.5}$$

From equation (2.4)

$$y^{(\alpha)}(t) = f^{\alpha}h(x(t)) + u(t) g f^{\alpha-1} h(x(t)), \tag{2.6}$$

where $\alpha = \alpha_h$. Thus, if y_d is any C^{∞} function that satisfies (2.5) and

$$u(x, t) = \frac{y_d^{(\alpha)}(t) - f^{\alpha}h(x)}{g f^{\alpha-1} h(x)}, \tag{2.7}$$

then

$$\begin{aligned} y^{(\alpha)}(t) &= f^{\alpha}h(x(t)) + u(x(t), t) g f^{\alpha-1} h(x(t)) \\ &= y_d^{(\alpha)}(t). \end{aligned}$$

Since $y^{\alpha}(t_0) = y_d^{\alpha}(t_0)$ it follows that $y \equiv y_d$. In particular, $y^{\alpha}(t_1) = y_d^{\alpha}(t_1) = h^{\alpha}(x_1)$ so that $x(t_1) = x_1$, as required.

This is a generalization of the technique used at the beginning of this section to transfer states for linear systems, where $g f^{\alpha-1} h(x) = cA^{\alpha-1}b$, a nonzero constant. Here $g f^{\alpha-1} h(x)$ can vanish.

DEFINITION. A state x is a *singular point for an output* $y = h(x)$ if $g f^{\alpha-1} h(x) = 0$. Let M_h^s denote the set of singular points for h in M.

The following results examine the behaviour of the system around singular points.

LEMMA 2.3. *Suppose that the system* (2.3) *is locally feedback linearizable and can be globally observed using the output* $y = h(x)$. *Then there exist* C^{∞} *functions* a_k, b_k : Range $h^{\alpha} \to R$ *with the following property*:

If $x(t)$ *is a trajectory then* $f^{\alpha}h(x(t)) = a_0(y^{\alpha}(t))$, $g f^{\alpha-1} h(x(t)) = b_0(y^{\alpha}(t))$,

and for each time t_s such that $x(t_s) \in M_h^s$ (i.e., $x(t_s)$ is a singular point),

$$\frac{d^k}{dt^k} f^\alpha h(x(t_s)) = a_k(y^\alpha(t_s)),$$

$$\frac{d^k}{dt^k} g f^{\alpha-1} h(x(t_s)) = b_k(y^\alpha(t_s))$$

for $k = 0, 1, 2, \cdots$ where $\alpha = \alpha_h$.

In particular, $x(t_s) \in M_h^s$ if and only if $b_0(h^\alpha(x(t_s))) = b_0(y^\alpha(t_s)) = 0$.

Proof. By definition $y^\alpha(t) = h^\alpha(x(t))$ so that $x(t) = (h^\alpha)^{-1}(y^\alpha(t))$ and thus $f^\alpha h(x(t)) = f^\alpha h((h^\alpha)^{-1}(y^\alpha(t))) = a_0(y^\alpha(t))$ where $a_0(z) = f^\alpha h((h^\alpha)^{-1}(z))$.

Similarly, if $b_0(z) = g f^{\alpha-1} h((h^\alpha)^{-1}(z))$ then $g f^{\alpha-1} h(x(t)) = b_0(y^\alpha(t))$. Now setting $\alpha = \alpha_h$, we have

$$\begin{aligned}
\frac{d}{dt} f^\alpha h(x)(t)) &= \frac{d}{dt} a_0(y^\alpha(t)) \\
&= \frac{d}{dt} a_0(y(t), y^{(1)}(t), \cdots, y^{(\alpha-1)}(t)) \\
&= \sum_{i=0}^{\alpha-1} \frac{\partial a_0}{\partial y^{(i)}} y^{(i+1)}(t) \\
&= (da_0)_{y^\alpha(t)}(y^{(1)}(t), y^{(2)}(t), \cdots, y^{(\alpha-1)}(t), y^{(\alpha)}(t)) \\
&= (da_0)_{y^\alpha(t)}(y^{(1)}(t), \cdots, y^{(\alpha-1)}(t), a_0(y^\alpha(t)) + u(t) b_0(y^\alpha(t))),
\end{aligned}$$

and at time t_s, $b_0(y^\alpha(t_s)) = 0$ so that

$$\frac{d}{dt} f^\alpha h(x(t_s)) = a_1(y^\alpha(t_s)) \quad \text{where } a_1(z) = (da_0)_z(z_1, \cdots, z_\alpha, a_0(z)).$$

Similarly, $d/dt\, g f^{\alpha-1} h(x(t_s)) = b_1(y^\alpha(t_s))$ and a simple induction argument completes the proof.

COROLLARY. *Suppose that the system* (2.3) *satisfies the hypotheses of Lemma* 2.3. *Then the functions a_k, b_k are defined inductively by*

$$a_{k+1}(z) = (da_k)_z(z_2, \cdots, z_\alpha, a_0(z)),$$

$$b_{k+1}(z) = (db_k)_z(z_2, \cdots, z_\alpha, a_0(z))$$

where $\alpha = \alpha_h$ and

$$a_0(z) = f^\alpha h((h^\alpha)^{-1}(z)),$$

$$b_0(z) = g f^{\alpha-1} h((h^\alpha)^{-1}(z))$$

for $z \in$ Range $h^\alpha \subseteq R^\alpha$.

Proof. These are the functions constructed in the proof of Lemma 2.3.

As in [12] we can define the *degree of singularity* $\beta(x_0)$ of a singular point x_0 that is connected with the number of extra restrictions placed on $y(t)$ when a trajectory

$x(t)$ passes through x_0 at time t. When (2.3) can be globally observed using the output $y=h(x)$, then $\beta(x)$ is simply the least positive integer k such that $b_k(h^\alpha(x_0))\neq 0$ and $\beta(x_0)=\infty$ if $b_k(h^\alpha(x_0))=0$ for all k. Note that if $x_0\notin M_h^s$ then $\beta(x_0)=0$.

LEMMA 2.4. *Suppose that the system* (2.3) *can be globally observed using the output* $y=h(x)$, *and* M, f, g, h *are real analytic. Then* $\beta(x_0)<\infty$ *for all* $x_0\in M$.

Proof. Suppose $\beta(x_0)=\infty$ for some $x_0\in M_h^s$. Then $b_k(h^\alpha(x_0))=0$ for all $k>0$. If $x(t)$ is any trajectory for (2.3) with $x(t_0)=x_0$, then from Lemma 2.3,

$$\frac{d^k}{dt^k} gf^{\alpha-1}h(x(t_0)) = b_k(y^\alpha(t_0)) = b_k(h^\alpha(x(t_0))) = 0$$

for $k=0, 1, 2, \cdots$, where $\alpha=\alpha_h$. Thus the Taylor coefficients in the Taylor series expansion of $gf^{\alpha-1}h(x(t))$ about $t=t_0$ are zero and hence $gf^{\alpha-1}h(x(t))=0$ for all $t\in R$. This, in turn, implies that $gf^{\alpha-1}h\equiv 0$ on M, which contradicts the definition of α (i.e., $gf^{\alpha-1}h\not\equiv 0$). Thus $\beta(x_0)<\infty$ and the proof is complete.

In light of Lemma 2.4 we can reasonably assume that $\beta(x)<\infty$ for all $x\in M$. If, in addition, the system (2.3) can be globally observed using the output $y=h(x)$, then the degree of a singularity can be expressed in terms of the outputs. That is, define

$$\hat{\beta}(z)=\beta((h^\alpha)^{-1}(z)),$$

a function defined on Range h^α. Thus when $x(t)$ is a trajectory for (2.3) with output $y(t)$ then $\beta(x(t))=\beta((h^\alpha)^{-1}h^\alpha(x(t)))=\hat{\beta}(y^\alpha(t))$, Now we can associate with this system the functions $a_0, a_1, \cdots, b_0, b_1, \cdots$ from Lemma 2.3 as well as the function $\hat{\beta}$.

DEFINITION. Let $\mathcal{Y}_h(x_0)$ denote the set of all C^∞ functions on R with the following properties:

(i) $y^\alpha(t)\in$ Range h^α for all t;

(ii) $y^\alpha(t_0)=h^\alpha(x_0)$;

(iii) If $b_0(y^\alpha(t_s))=0$ at time t_s, then $y^{(\alpha_h+k)}(t_s)=a_k(y^\alpha(t_s))$ for $0\leqq k<\hat{\beta}(y^\alpha(t_s))$.

THEOREM 2.5. *Suppose that the system* (2.3) *can be globally observed using the output* $y=h(x)$. *Then each function in* $\mathcal{Y}_h(x_0)$ *is an output for the system resulting from a continuous input. Furthermore,* $\mathcal{Y}_h(x_0)$ *contains all output functions which result from* C^∞ *inputs.*

Proof. Let u be a C^∞ input to the system (2.3) that results in the trajectory $x(t)$ and the output function $y(t)=h(x(t))$. Then $y^\alpha(t)=h^\alpha(x(t))$, and thus $y^\alpha(t)\in$ Range h^α for $t\geqq t_0$. Also $y^\alpha(t_0)=h^\alpha(x(t_0))=h^\alpha(x_0)$. Finally, suppose that $b_0(y^\alpha(t_s))=0$ for some time t_s. Let $\alpha=\alpha_h$. From Lemma 2.3 $y^{(\alpha)}(t)=a_0(y^\alpha(t))+u(t)b_0(y^\alpha(t))$. Thus if $\beta=\hat{\beta}(y^\alpha(t_s))$ then $b_0(y^\alpha(t_s))=(d/dt)b_0(y^\alpha(t_s))=\cdots=(d^{\beta-1}/dt^{\beta-1})b_0(y^\alpha(t_s))=0$ and $(d^\beta/dt^\beta)b_0(y^\alpha(t_s))\neq 0$. This means that

$$y^{(\alpha+1)}(t)=\frac{d}{dt}a_0(y^\alpha(t))+u^{(1)}(t)b_0(y^\alpha(t))+u(t)\frac{d}{dt}b_0(y^\alpha(t))$$

so at time t_s

$$y^{(\alpha+1)}(t_s)=\frac{d}{dt}a_0(y^\alpha(t_s))+u^{(1)}(t_s)\cdot 0+u(t_s)\cdot 0$$

$$=a_1(y^\alpha(t_s)).$$

Similarly, $y^{(\alpha+k)}(t_s)=a_k(y^\alpha(t_s))$ for $0\leqq k<\beta$. Thus if u is C^∞ then $y\in\mathcal{Y}_h(x_0)$.

Finally, suppose that $y_d \in \mathcal{Y}_h(x_0)$. Define the feedback control

$$u_d(t,x)=\begin{cases}\dfrac{y_d^{(\alpha)}(t)-a_0(h^\alpha(x))}{b_0(h^\alpha(x))} & \text{when } b_0(y_d^{(\alpha)}(t_s))\neq 0,\\ r(t_s) \quad \text{when } b_0(y_d^\alpha(t_s))=0\end{cases}$$

where $r(t_s)=(y_d^{(\alpha+\beta)}(t_s)-a_\beta(y_d^\alpha(t_s)))/b_\beta(y_d^\alpha(t_s))$. The proof will be complete if it can be shown that using u_d results in a trajectory $x(t)$ with the property that $t\mapsto u_d(t,x(t))$ is continuous and $y(t)=h(x(t))=y_d(t)$. Of course $y_d^\alpha(t_0)=h^\alpha(x_0)=y^\alpha(t_0)$ so that $y\equiv y_d$ if $y^{(\alpha)}\equiv y_d^{(\alpha)}$. If $t\neq t_s$, then

$$\begin{aligned}y^{(\alpha)}(t)&=a_0(y^\alpha(t))+u_d(t,x(t))b_0(y^\alpha(t))\\&=a_0(y^\alpha(t))+\frac{y_d^{(\alpha)}(t)-a_0(y^\alpha(t))}{b_0(y^\alpha(t))}\cdot b_0(y^\alpha(t))\\&=y_d^{(\alpha)}(t),\end{aligned}$$

so that $y\equiv y_d$ for $t_0\leqq t<t_s$. At time t_s, $b_0(y_d^\alpha(t_s))=0$ and $u_d(t_s,x(t_s))=r(t_s)$. Now

$$\begin{aligned}\lim_{t\to t_s^-}u_d(t,x(t))&=\lim_{t\to t_s^-}\frac{y_d^{(\alpha)}(t)-a_0(y_d^\alpha(t))}{b_0(y_d^\alpha(t))}\\&=\lim_{t\to t_s^-}\frac{y_d^{(\alpha+1)}(t)-(d/dt)a_0(y_d^\alpha(t))}{(d/dt)b_0(y_d^\alpha(t))}\end{aligned}$$

using l'Hôpital's rule—note that $y_d^{(\alpha)}(t_s)=a_0(y_d^\alpha(t_s))$ since $y\in\mathcal{Y}_h(x_0)$. Now if $\beta>1$, then $(d/dt)b_0(y_d^\alpha(t_s))=b_1(y_d^\alpha(t_s))=0$ and $y_d^{(\alpha+1)}(t_s)=(d/dt)a_0(y_d^\alpha(t_s))=a_1(y_d^\alpha(t_s))$ since $y\in\mathcal{Y}_h(x_0)$, so using l'Hôpital's rule

$$\lim_{t\to t_s^-}u_d(t,x(t))=\lim_{t\to t_s^-}\frac{y_d^{(\alpha+2)}(t)-(d^2/dt^2)a_0(y_d^\alpha(t))}{(d^2/dt^2)b_0(y_d^\alpha(t))}$$

and continuing this process

$$\begin{aligned}\lim_{t\to t_s^-}u_d(t,x(t))&=\lim_{t\to t_s}\frac{y_d^{(\alpha+\beta)}(t)-(d^\beta/dt^\beta)a_0(y_d^\alpha(t))}{(d^\beta/dt^\beta)b_0(y_d^\alpha(t))}\\&=\frac{y_d^{(\alpha+\beta)}(t_s)-a_\beta(y_d^\alpha(t_s))}{b_\beta(y_d^\alpha(t_s))}\\&=r(t_s)\\&=u_d(t_s,x(t_s))\end{aligned}$$

and it follows that u_d is continuous and the proof is complete. Note that in open-loop form the control corresponding to y_d is

$$u_d(t)=\begin{cases}\dfrac{y_d^{(\alpha)}(t)-a_0(y_d^\alpha(t))}{b_0(y_d^\alpha(t))} & \text{when } b_0(y_d^\alpha(t))\neq 0,\\ r(t_s) \quad \text{when } b_0(y_d^\alpha(t_s))=0\end{cases}\tag{2.8}$$

where $r(t_s)=(y_d^{(\alpha+\beta)}(t_s)-a_\beta(y_d^\alpha(t_s)))/b_\beta(y_d^\alpha(t_s))$.

In § 3 the set $\mathcal{Y}_h(x_0)$ of output functions is used to deduce the reachable set and perform state transfers using inputs derived from output tracking. This section ends with a number of examples.

Example 2.3 (Example 1.1 continued). Here $M = R^2$, $f(x) = (0, e^{x_1})$, $g(x) = (1, 0)$, and $x(0) = x_0 = (a, b)$. To find an output $h(x)$ with $\alpha_h = \dim M = 2$ requires $gh \equiv 0$. Thus h is a function of x_2. We are led to try $h(x_1, x_2) = x_2$. Thus $gh \equiv 0$, $gfh(x_1, x_2) = e^{x_1}$ so that $\alpha_h = 2$, and since $h^\alpha(x_1, x_2) = (x_2, e^{x_1})$, a diffeomorphism of M into R^2, this system can be globally observed using the output $y = h(x)$. Since $gf^{\alpha-1}(x_1, x_2) = e^{x_1} \neq 0$ there are no singular points, i.e., $\beta(x) = 0$ for all $x \in M$, and thus $\mathcal{Y}_h(x_0)$ consists of C^∞ functions y with

(i) $(y(t), \dot{y}(t)) \in \text{Range } h^\alpha = R \times R^+ \qquad (R^+ = (0, \infty))$,

(ii) $(y(0), \dot{y}(0)) = h^\alpha(x_0) = (b, e^a)$.

That is, $\mathcal{Y}_h(x_0) = \{y \in C^\infty(R) \mid y(0) = b,\ \dot{y}(0) = e^a \text{ and } y \text{ increasing}\}$.

Example 2.4. Consider the system model where $f(x_1, x_2) = (0, x_1 + x_2 + 1)$, $g(x_1, x_2) = (x_1 + x_2, 0)$ and $x(0) = x_0 = (a, b) \in M = R^2$. Here $\text{ad}_f g(x_1, x_2) = (x_1 + x_2 + 1, -(x_1 + x_2))$ so that condition (2.4) is satisfied (when $a + b \neq 0$). To find an output $h : M \to R^2$ such that the system is globally observed we require $\alpha_h = \dim M = 2$. Since this means $gh = (\partial h/\partial x_1)(x_1 + x_2) \equiv 0$, h must be a function of x_2 alone. Trying $h(x_1, x_2) = x_2$ we have $fh(x_1, x_2) = x_1 + x_2 + 1$ so that $h^\alpha(x_1, x_2) = (x_2, x_1 + x_2 + 1)$ is a diffeomorphism of $M = R^2$ into R^2 (in fact, Range $h^\alpha = R^2$), and h will suffice. Since

$$\begin{aligned} y &= h(x_1, x_2) = x_2, \\ \dot{y} &= fh(x_1, x_2) = x_1 + x_2 + 1, \\ \ddot{y} &= f^2 h(x_1, x_2) + ugfh(x_1, x_2) \\ &= (x_1 + x_2 + 1) + u(x_1 + x_2) \\ &= \dot{y} + u(\dot{y} - 1) \end{aligned}$$

it follows that $a_0(y, \dot{y}) = \dot{y}$ and $b_0(y, \dot{y}) = \dot{y} - 1$. Here the singular set is $M_h^s = \{(x_1, x_2) \mid x_1 + x_2 = 0\}$, and from the corollary to Lemma 2.3,

$$\begin{aligned} a_1(y, \dot{y}) &= da_{0_{(y,\dot{y})}}(\dot{y}, \dot{y}) = \dot{y} \\ b_1(y, \dot{y}) &= db_{0_{(y,\dot{y})}}(\dot{y}, \dot{y}) = \dot{y}. \end{aligned}$$

Thus if $b_0(y^\alpha(t_s)) = b_0(y(t_s), \dot{y}(t_s)) = \dot{y}(t_s) - 1 = 0$ at time t_s (i.e., $\dot{y}(t_s) = 1$), then $b_1(y^\alpha(t_s)) = \dot{y}(t_s) = 1 \neq 0$, hence $\beta = 1$. Since Range $h^\alpha = R^2$, it follows that $\mathcal{Y}_h(x_0)$ consists of all C^∞ functions with the property that

$y^\alpha(0) = (y(0), \dot{y}(0)) = h^\alpha(x_0) = (b, a + b + 1)$ and if $b_0(y^\alpha(t_s)) = \dot{y}(t_s) - 1 = 0$ then $\ddot{y}(t_s) = a_1(y^\alpha(t_s)) = \dot{y}(t_s) = 1$.

That is, $\mathcal{Y}_h(x_0) = \{y \in C^\infty(R) \mid y(0) = b,\ \dot{y}(0) = a + b + 1$, and when $\dot{y}(t_2) = 1$ then $\ddot{y}(t_s) = 1\}$.

Example 2.5. Consider the system

$$\begin{aligned} \dot{x}_1 &= x_1 u, & x_1(0) &= a, \\ \dot{x}_2 &= \ln x_1, & x_2(0) &= b \end{aligned}$$

where $M = R^+ \times R$. This system can be globally observed using $h(x) = x_2$ (or e^{x_2}, etc.). Here $h^\alpha(x_1, x_2) = (x_2, \ln x_1) : M \to R^2$ is a diffeomorphism onto R^2, and $\ddot{y} = u$ so that $gfh(x_1, x_2) = 1$ and there are *no* singular states. Thus

$$\mathcal{Y}_h(x_0) = \{y \in C^\infty(R) \mid y(0) = b,\ \dot{y}(0) = \ln a\}.$$

3. The set of reachable states. Let $x(t, u, x_0)$ denote the solution to the differential equation (2.3) where u is a continuous control. A state $x \in M$ is said to be *reachable*

from x_0 at time t if $x = x(t, y, x_0)$ for some (continuous) control u. The collection of all states reachable from x_0 at time t is denoted by $\mathcal{R}_t(x_0)$, the *reachable set at time t.* The system (2.3) is called *strongly controllable* if $\mathcal{R}_t(x_0) = M$ for all $t > t_0$.

For systems that can be globally observed using an output h we can essentially determine $\mathcal{R}_t(x_0)$ from the collection of output functions $\mathcal{Y}_h(x_0)$. Define $\mathcal{Q}_t^h(x_0)$ to be the subset

$$\mathcal{Q}_t^h(x_0) = \{(h^\alpha)^{-1} y^\alpha(t) \,|\, y \in \mathcal{Y}_h(x_0)\} \subseteq M.$$

THEOREM 3.1. *Suppose that the system* (2.3) *can be globally observed using the output $y = h(x)$. Then*

(i) $\mathcal{R}_t(x_0) \supseteq \mathcal{Q}_t^h(x_0)$ *for all* $t > t_0$;

(ii) $\mathcal{Q}_t^h(x_0)$ *is dense in* $\mathcal{R}_t(x_0)$, *in fact* $\operatorname{cl}(\operatorname{int} \mathcal{Q}_t^h(x_0)) = \operatorname{cl}(\operatorname{int} \mathcal{R}_t(x_0))$ *for all* $t > t_0$; *and*

(iii) $\mathcal{R}_t(x_0) = M$ *if and only if* $\mathcal{Q}_t^h(x_0) = M$.

COROLLARY. *Suppose that the system* (2.3) *can be globally observed using the output $y = h(x)$. Then we can transfer the state of the system from $x(t_0) = x_0$ to $x(t_1) = x_1 \in \mathcal{Q}_t^h(x_0)$ as follows*:

(i) *Choose $y \in \mathcal{Y}_h(x_0)$ such that $y(t_1) = h^\alpha(x_1)$.*

(ii) *Set $u_d(t) = (y^{(\alpha)}(t) - a_0(y^\alpha(t)))/b_0(y^\alpha(t))$ for all t such that $b_0(y^\alpha(t)) \neq 0$ $(\alpha = \alpha_h)$.*

(iii) *If $b_0(y^\alpha(t_s)) = 0$ set $u_d(t_s) = (y^{(\alpha+\beta)}(t_s) - a_\beta(y^\alpha(t_s)))/b_\beta(y^\alpha(t_s))$ where $\beta = \hat{\beta}(y^\alpha(t_s))$.*

Then u_d is a continuous control that steers the system from $x(t_0) = x_0$ to $x(t_1) = x_1$.

Proof of Theorem 3.1. From Theorem 2.5 $\mathcal{Y}_h(x_0)$ contains all outputs that result from C^∞ inputs, and each $y \in \mathcal{Y}_h(x_0)$ is an output $y(t) = h(x(t))$ for the system corresponding to a continuous input $u(t)$. Thus $y^\alpha(t) = h^\alpha(x(t))$ and $(h^\alpha)^{-1} y^\alpha(\cdot) = x(\cdot)$ is a system trajectory so that (i) follows. Since the set of states we can reach at time t using C^∞ inputs has an interior that is dense in $\mathcal{R}_t(x_0)$ (cf. [15]), (ii) follows. Of course $\mathcal{R}_t(x_0) = M$ if and only if $\operatorname{int} \mathcal{R}_t(x_0) = M$, if and only if $\mathcal{Q}_t^h(x_0) = M$, and the proof is complete.

Proof of the corollary. This result is proved in § 2. In particular, $u(t)$ is described in (2.8).

Example 3.1. Consider the system from Example 2.3. Here $y \in \mathcal{Y}_h(x_0)$ if and only if $y(0) = b$, $\dot{y}(0) = e^a$ and y is increasing $(x(0) = (a, b))$. Thus $y(t) > b$ and $\dot{y}(t) > 0$ for any $t > 0$, so that

$$\mathcal{Q}_t^h(x_0) = (h^\alpha)^{-1}\{(y_1, y_2) \,|\, y_1 > b, y_2 > 0\}.$$

Furthermore, $h^\alpha(x_1, x_2) = (x_2, e^{x_1})$ so that $(h^\alpha)^{-1}(y_1, y_2) = (\ln y_2, y_1)$, and thus

$$\begin{aligned} \mathcal{Q}_t^h(x_0) &= \{(\ln y_2, y_1) \,|\, y_1 > b, y_2 > 0\} \\ &= R \times (b, \infty). \end{aligned}$$

Since $\mathcal{Q}_t^h(x_0)$ is a dense subset of $\mathcal{R}_t(x_0)$, and $M = R^2$, clearly the system is not strongly controllable. In fact,

$$\operatorname{cl} \mathcal{R}_t(x_0) = \operatorname{cl} \mathcal{Q}_t^h(x_0) = R \times [b, \infty) \neq R^2.$$

Example 3.2. Consider the system from Example 2.4. Here $M = R^2$, $h^\alpha(x_1, x_2) = (x_2, x_1 + x_2 + 1)$ so that $(h^\alpha)^{-1}(y_1, y_2) = (y_2 - y_1 - 1, y_1)$, and $x(0) = (a, b)$. Suppose $x(0) = (1, 1)$. Then $y \in \mathcal{Y}_h(x_0)$ if and only if $y(0) = 1$, $\dot{y}(0) = 3$ and whenever $\dot{y}(t_s) = 1$

then $\ddot{y}(t_s)=1$. This means that $\dot{y}$ is increasing when $\dot{y}(t_s)=1$, and hence if $y\in\mathcal{Y}_h(x_0)$ then $\dot{y}(t)$ will always be greater than one. Thus $y(t)>y(0)+t=t+1$ and

$$\begin{aligned}\mathcal{Q}_t^h(x_0)&=(h^\alpha)^{-1}\{(y_1,y_2)\,|\,y_1>t+1,\,y_2>1\}\\&=\{(y_2-y_1-1,y_1)\,|\,y_1>t+1,\,y_2>1\}\\&=\{(x_1,x_2)\,|\,x_2>t+1,\,x_1>-x_2\}.\end{aligned}$$

Thus $\operatorname{cl}\mathcal{R}_t(x_0)=\operatorname{cl}\mathcal{Q}_t^h(x_0)=\{x_2\geqq t+1,\,x_1\geqq -x_2\}\neq R^2=M$, and the system is far from being strongly controllable.

Example 3.3. From Example 2.5, $y\in\mathcal{Y}_h(x_0)$ if and only if $y(0)=b$, $\dot{y}(0)=\ln a$ where $x(0)=(a,b)$ so that $y(t)$ and $\dot{y}(t)$ are unrestricted. Here $h^\alpha(x_1,x_2)=(x_2,\ln x_1)$ so that $(h^\alpha)^{-1}(y_1,y_2)=(e^{y_2},y_1)$ and

$$\begin{aligned}Q_t^h(x_0)&=(h^\alpha)^{-1}(R^2)\\&=M.\end{aligned}$$

Thus $\mathcal{R}_t(x_0)=M$ and the system is strongly controllable.

In Example 3.1 there are no singular states but the range of $h^\alpha\neq R^2$ and the system is not strongly controllable. In Example 3.1 Range $h^\alpha=R^2$ but there are singular states, i.e., $M_h^s=\varnothing$, and this system also failed to be strongly controllable. Example 3.3 is a strongly controllable system with $M_h^s=\varnothing$ and Range $h^\alpha=R^2$. The connections between $\mathcal{R}_t(x_0)$ and Range h^α are not explored, but when Range $h^\alpha=R^n$ ($n=\dim M$) strong controllability is related to the existence of singular states.

THEOREM 3.2. *Suppose that the system* (2.3) *can be globally observed using the output* $y=h(x)$ *with* Range $h^\alpha=R^n$. *If* $M_h^s=\varnothing$ *then the system* (2.3) *is strongly controllable. If* $M_h^s\neq\varnothing$ *and* $b_1(h^\alpha(x))>0$ (*or* <0) *on* M_h^s (*e.g.*, $\beta=1$ *on* M_h^s *and* M_h^s *is connected*) *then* (2.3) *is not strongly controllable.*

Proof. Suppose that the above system has $M_h^s=\varnothing$. Then

$$\begin{aligned}&\mathcal{Y}_h(x_0)=\{y\in C^\infty(R)\,|\,y^\alpha(t_0)=h^\alpha(x_0)\},\quad\text{and}\\&\mathcal{Q}_t^h(x_0)=\{(h^\alpha)^{-1}y^\alpha(t)\,|\,y\in\mathcal{Y}_h(x_0)\}\\&\qquad=\{(h^\alpha)^{-1}z\,|\,z\in R^n\}\\&\qquad=M.\end{aligned}$$

Thus using Theorem 3.1 we conclude that $\mathcal{R}_t(x_0)=M$ for any $x_0\in M$ and $t>t_0$.

Now suppose $b_1(h^\alpha(x))>0$ on M_h^s, $M_h^s\neq\varnothing$, and the system (2.3) is strongly controllable. Choose $x_1\in M_h^s$. Then there exist a control u and trajectory $t\mapsto x(t)$ such that $x(t_0)=x_0$, $x(t_s)=x_1$ and $x(2t_s)=x_1$ where $t_s>t_0$. Here $x(t_s)\in M_h^s$ so that

$$b_0(h^\alpha(x(t_s)))=b_0(y^\alpha(t_s))=0$$

$$\frac{d}{dt}b_0(h^\alpha(x(t_s))=\frac{d}{dt}b_0(y^\alpha(t_s))=b_1(y^\alpha(t_s))>0$$

by assumption. This means that $b_0(y^\alpha(t))$ is increasing at time t_s. Also $x(2t_s)\in M_h^s$ so that $b_0(y^\alpha(2t_s))=0$, and hence $b_0(y^\alpha(t))$ must become zero on $(t_s,2t_s]$. If $b_0(y^\alpha(t_1))=0$ for some $t_1\in(t_s,2t_s]$ then $(d/dt)b_0(y^\alpha(t_1))\leqq 0$, which contradicts $b_1(h^\alpha(x))>0$ for all $x\in M_h^s$. Thus (2.3) cannot be strongly controllable.

Acknowledgment. The author thanks the anonymous referee for constructive criticism and helpful suggestions.

REFERENCES

[1] R. W. BROCKETT, *Feedback invariants for nonlinear systems*, IFAC Congress, Helsinki, Finland, 1978.
[2] B. JAKUBCZYK AND W. RESPONDEK, *On linearization of control systems*, Bull. Acad. Polon. Sci., Sér. Sci. Math. Astronom. Phys., 28 (1980), pp. 517-522.
[3] L. R. HUNT AND R. SU, *Linear equivalents of nonlinear time varying systems*, International Symposium on the Mathematical Theory of Networks and Systems, Santa Monica, CA, 1981, pp. 119-123.
[4] A. ISIDORI AND A. KRENNER, *On the feedback equivalence of nonlinear systems*, Systems Control Lett., 2 (1982), pp 118-121.
[5] A. ISIDORI AND A. RUBERTI, *On the synthesis of linear input-output responses for nonlinear systems*, Systems Control Lett., 4 (1985), pp. 17-22.
[6] R. MARINO, W. M. BOOTHBY, AND D. L. ELLIOT, *Geometric properties of linearizable control systems*, Math. Systems Theory, 18 (1985) pp. 97-123.
[7] M. ILIC-SPONG, R. MARINO, S. M. PERESADA, AND D. G. TAYLOR, *Feedback linearizable control of switched reluctance motors*, IEEE Trans. Automat. Control, 32 (1987), pp. 371-380.
[8] S. N. SINGH AND A. A. SCHY, *Invertibility and robust nonlinear control of robotic systems*, in Proc. 23rd IEEE Conference on Decision and Control, Las Vegas, NV, 1984.
[9] A. DE LUCA, A. ISIDORI, AND F. NICOLO, *An application of nonlinear model matching to the dynamic control of a robot arm with elastic joints*, in Proc. 1st IFAC Symposium Robot Control, Barcelona, Spain, 1985.
[10] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
[11] R. HIRSCHORN, *Invertibility of nonlinear control systems*, SIAM J. Control Optim. 17 (1979), pp. 289-297.
[12] R. HIRSCHORN AND J. DAVIS, *Output tracking of nonlinear systems with singular points*, SIAM J. Control Optim., 25 (1987), pp. 547-557.
[13] ———, *Global output tracking for nonlinear systems*, SIAM J. Control Optim., 26 (1988), pp. 1321-1330.
[14] F. WARNER, *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman, Glenview, IL, 1970.
[15] H. SUSSMAN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95-116.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 552-569, May 1990

005

AN APPROXIMATION THEORY FOR THE IDENTIFICATION OF NONLINEAR DISTRIBUTED PARAMETER SYSTEMS*

H. T. BANKS†, SIMEON REICH‡, AND I. G. ROSEN§

Abstract. An abstract approximation framework for the identification of nonlinear distributed parameter systems is developed. Inverse problems for nonlinear systems governed by strongly maximal monotone operators (satisfying a mild continuous dependence condition with respect to the unknown parameters to be identified) are treated. Convergence of Galerkin approximations and the corresponding solutions of finite-dimensional approximating identification problems to a solution of the original infinite-dimensional identification problem is demonstrated, using the theory of nonlinear evolution systems and a nonlinear analogue of the Trotter–Kato approximation result for semigroups of bounded linear operators. The nonlinear theory developed here is shown to subsume an existing linear theory as a special case. It is also shown to be applicable to a broad class of nonlinear elliptic operators and the corresponding nonlinear parabolic partial differential equations to which they lead. An application of the theory to a quasilinear model for heat conduction or mass transfer is discussed.

Key words. nonlinear evolution systems, nonlinear distributed parameter systems, maximal monotone operator, identification, Galerkin approximation, nonlinear heat conduction

AMS(MOS) subject classifications. 47H20, 93C10, 93C20

1. Introduction.

In this paper we develop a general abstract approximation framework for the identification of nonlinear distributed parameter evolution systems. Our intent is to define relatively straightforward and easily verified criteria that are applicable to broad classes of nonlinear systems; these criteria will guarantee the convergence of solutions to a sequence of finite-dimensional Galerkin approximation based parameter estimation problems to a solution of the original, underlying, infinite-dimensional identification problem. The results that we present below generalize and extend the theory recently developed by Banks and Ito in [2] and [3] for regularly dissipative or abstract parabolic, linear systems. It is, to the best of our knowledge, the first such general approximation theory for inverse problems involving nonlinear distributed systems.

The sufficient conditions set down in our framework include a relatively weak continuity assumption with respect to the unknown parameters to be identified, an equiboundedness and an equistrong monotonicity assumption on the nonlinear operator describing the system dynamics. In addition our theory requires a standard approximation assumption on the Galerkin subspaces used to effect the finite-dimensional, or finite-element, approximations. We demonstrate that solutions to the

* Received by the editors April 20, 1988; accepted for publication (in revised form) November 11, 1988. Part of this research was carried out while the first and third authors were visiting scientists at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, Virginia, which is operated under National Aeronautics and Space Administration contract NAS1-18107.

† Center for Control Sciences, Division of Applied Mathematics, Brown University, Providence, Rhode Island. The research of this author was supported in part by National Science Foundation grant MCS-8504316, National Aeronautics and Space Administration grant NAG-1-517, and Air Force Office of Scientific Research grants AFOSR-84-0398 and AFOSR-F49620-86-C-0111.

‡ Department of Mathematics, University of Southern California, Los Angeles, California and Department of Mathematics, The Technion, Israel Institute of Technology, Haifa, Israel. The research of this author was supported in part by the Fund for the Promotion of Research at The Technion and by the Technion Vice Presidents Research Fund.

§ Department of Mathematics, University of Southern California, Los Angeles, California. The research of this author was supported in part by Air Force Office of Scientific Research grants AFOSR-84-0393 and AFOSR-87-0356.

finite-dimensional identification problems approximate a solution to the infinite-dimensional identification problem via a convergence result for solutions to the forward problems. This result is obtained using the theory of nonlinear evolution systems and a nonlinear analog of the well-known Trotter–Kato approximation result for linear semigroups.

In the present paper, we are concerned only with theory; implementation questions and conclusions drawn from our numerical or computational studies will be reported on elsewhere. Also, while we have tried to make our framework as versatile as possible, the treatment below does have limitations. For example, our theory can handle quasi-autonomous systems but it is not applicable in the fully nonautonomous case. The development of a general theory that can handle nonlinear systems involving time-dependent operators requires additional effort and is currently the focus of our ongoing investigations. The particular difficulties inherent in the time-dependent case will be described in greater detail in our discussions below.

We provide a brief outline of the remainder of the paper. In § 2 we state a fundamental existence and uniqueness result for infinite-dimensional nonlinear systems and prove a general approximation result that is especially well suited for application in the context of the inverse problems which are the central focus of our study. In § 3 we define a class of nonlinear distributed systems and the associated parameter identification problems. We define the Galerkin approximations and prove the general convergence result. Section 4 contains some examples. We show that our nonlinear theory subsumes the linear theory presented in [2] and [3] as a special case; we also consider the application of our framework to a class of nonlinear elliptic operators and the corresponding nonlinear parabolic partial differential equations to which they lead. In particular, we look at the application of our results to a well-known quasilinear model for heat conduction or mass transfer. In § 5 we summarize our findings and provide some concluding remarks.

2. An approximation result for nonlinear evolution systems. Let X_0 be a Banach space with norm $|\cdot|_0$. We consider the nonlinear, quasi-autonomous initial value problem in X_0 given by

$$\dot{x}_0(t) + A_0 x_0(t) \ni f_0(t), \qquad 0 < t \leqq T, \tag{2.1}$$

$$x_0(0) = x_0^0 \tag{2.2}$$

where $x_0^0 \in X_0$, $f_0 : [0, T] \to X_0$ and the nonlinear operator $A_0 : X_0 \to 2^{X_0}$ is in general multivalued, not everywhere defined, and not continuous. The existence of solutions to the initial value problem (2.1), (2.2) and the subsequent approximation result to follow, are both consequences of Theorem 2.1 to be given below.

We require the following definitions. Let X be a Banach space with norm $|\cdot|_X$. For $A : X \to 2^X$, a nonlinear, multivalued operator, the domain and range of A are defined by $\mathrm{Dom}(A) = \{x \in X : Ax \neq \emptyset\}$ and $\mathcal{R}(A) = \bigcup_{x \in \mathrm{dom}(A)} Ax$, respectively. We say that the operator A is accretive if for every $\lambda > 0$, $x_1, x_2 \in \mathrm{Dom}(A)$ and $y_1 \in Ax_1$, $y_2 \in Ax_2$, we have

$$|x_1 - x_2|_X \leqq |x_1 - x_2 + \lambda(y_1 - y_2)|_X.$$

We say that A is m-accretive if A is accretive and $\mathcal{R}(I + \lambda A) = X$ for some $\lambda > 0$. We note that if A is m-accretive, then $\mathcal{R}(I + \lambda A) = X$ for every $\lambda > 0$ and for each $\lambda > 0$ the resolvent of A at λ, $J(\lambda; A) : X \to X$, a single-valued, everywhere defined, nonlinear operator on X can be defined as $J(\lambda; A) = (I + \lambda A)^{-1}$.

A two-parameter family of nonlinear operators $\{U(t, s)\colon 0 \leqq s \leqq t \leqq T\}$ defined on a subset $\Omega \subset X$ is called a nonlinear evolution system on Ω if for each $x \in \Omega$ we have $U(t, s)x \in \Omega$, $U(s, s)x = x$, and $U(t, r)U(r, s)x = U(t, s)x$ for $0 \leqq s \leqq r \leqq t \leqq T$, and $U(t, s)x$ is continuous from the triangle $\triangle = \{[s, t]\colon 0 \leqq s \leqq t \leqq T\}$ into X.

A strongly continuous function $x\colon [0, T] \to X$ is called a strong solution to the quasi-autonomous initial value problem

$$\dot{x}(t) + Ax(t) \ni f(t), \qquad 0 < t \leqq T, \tag{2.3}$$

$$x(0) = x^0 \tag{2.4}$$

where $f\colon [0, T] \to X$ and $x^0 \in X$ if x is absolutely continuous on compact subintervals of $(0, T)$, differentiable almost everywhere and satisfies $f(t) - \dot{x}(t) \in Ax(t)$ for almost every $t \in [0, T]$ and $x(0) = x^0$.

THEOREM 2.1. *Let X be a Banach space with norm $|\cdot|_X$ and suppose that $A\colon X \to 2^X$ and $f\colon [0, T] \to X$ appearing in* (2.3) *satisfy*

(1) *That there exists an $\omega \in R$ for which the operator $A + \omega I$ is m-accretive,*

(2) *$f \in L_1(0, T; X)$.*

Then a unique, nonlinear evolution system $\{U(t, s)\colon 0 \leqq s \leqq t \leqq T\}$ on $\overline{\mathrm{Dom}(A)}$ can be constructed that satisfies

(i) $|U(t, s)\phi - U(t, s)\psi|_X \leqq e^{\omega(t-s)}|\phi - \psi|_X$, *for* $\phi, \psi \in \overline{\mathrm{Dom}(A)}$ *and* $0 \leqq s \leqq t \leqq T$,

(ii) $|U(s+t, s)\phi - U(r+t, r)\phi|_X \leqq 2\int_0^t e^{\omega(t-\tau)}|f(\tau+s) - f(\tau+r)|_X \, d\tau$, *for all* $\phi \in \overline{\mathrm{Dom}(A)}$ *and all* $t > 0$ *such that* $s + t$, $r + t \leqq T$.

(iii) *If* $x^0 \in \overline{\mathrm{Dom}(A)}$ *and the initial value problem* (2.3), (2.4) *has a strong solution x, then*

$$x(t) = U(t, s)x(s) \quad \text{for } 0 \leqq s \leqq t \leqq T.$$

When $x^0 \in \overline{\mathrm{Dom}(A)}$, the strongly continuous function $x\colon [0, T] \to X$ given by $x(t) = U(t, 0)x^0$ is referred to as a mild or generalized solution to (2.3), (2.4).

Theorem 2.1 is a direct consequence of results given by Crandall and Evans in [7] and [9]. Henceforth, we will assume that $A_0\colon X_0 \to 2^{X_0}$ and $f_0\colon [0, T] \to X$, satisfy (1) and (2) in the statement of Theorem 2.1 and that $x_0 \in \overline{\mathrm{Dom}(A_0)}$. We then let $\{U_0(t, s)\colon 0 \leqq s \leqq t \leqq T\}$ denote the corresponding nonlinear evolution system on $\overline{\mathrm{Dom}(A_0)}$ and consider the approximation of mild solutions to the initial value problem (2.1), (2.2).

Our approximation result is in the spirit of those given for nonlinear semigroups and evolution systems by Crandall and Pazy in [8] and Goldstein in [10]. However, our theorem differs from these earlier treatments in two ways. First, we require that the time-dependent perturbation f_0 be only L_1 as opposed to it being continuous as in [8] and satisfying a Lipschitz-like condition in [10]. This distinction is especially relevant in the case of control systems where discontinuous input is common. The second difference is that we give our result in a form that is most appropriate for application to the development of a general approximation theory or framework and computational schemes for the parameter identification problems to be discussed in the next section.

We require some set theoretic notation. For sets H_n, $n = 0, 1, 2, \cdots$, by $\lim H_n \supset H_0$ we mean: given $x_0 \in H_0$, there exist $x_n \in H_n$ such that $x_n \to x_0$ as $n \to \infty$.

THEOREM 2.2. *For each $n \in Z^+ = \{1, 2, 3, \cdots\}$ let X_n be a closed linear subspace of X_0. For $n = 0, 1, \cdots$, let $A_n\colon X_n \to 2^{X_n}$ be a possibly multivalued nonlinear operator on X_n, and let $f_n\colon [0, T] \to X_n$ be an X_n-valued measurable function defined on $[0, T]$. Suppose that there exists an $\omega_0 \in R$, independent of n, for which the operators $A_n + \omega_0 I$ are m-accretive, that there exists a function $g \in L_1(0, T; X_0)$ for which $|f_n(t)| \leqq g(t)$, for*

almost every $t \in [0, T]$, *and that* $\lim \bar{D}_n \supset \bar{D}_0$ *where* $D_n = \mathrm{Dom}(A_n)$ *and* $D_0 = \mathrm{Dom}(A_0)$. *Suppose further that for some* $\lambda_0 > 0$, *we have*

$$\lim_{n\to\infty} J(\lambda_0; A_n + \omega_0 I)\phi_n = J(\lambda_0; A + \omega_0 I)\phi_0 \tag{2.5}$$

whenever $\phi_n \in X_n$ *with* $\lim_{n\to\infty} \phi_n = \phi_0 \in X_0$, *and that*

$$\lim_{n\to\infty} f_n(t) = f_0(t) \quad \textit{for a.e. } t \in [0, T].$$

Then for each $n \in Z^+$ *there exists a unique nonlinear evolution system* $\{U_n(t, s): 0 \leqq s \leqq t \leqq T\}$ *on* $\bar{D}_n$ *corresponding (in the sense of Theorem* 2.1) *to* A_n *and* f_n *and for* $\phi_n \in \bar{D}_n$ *with* $\lim_{n\to\infty} \phi_n = \phi_0 \in \bar{D}_0$ *we have*

$$\lim_{n\to\infty} U_n(t, s)\phi_n = U_0(t, s)\phi_0, \qquad 0 \leqq s \leqq t \leqq T, \tag{2.6}$$

with the limit being uniform in t *for* $t \in [s, T]$.

Proof. We follow Goldstein (see [10], [11]) and use an approach first suggested by Kisynski [13] for demonstrating the convergence of approximations to linear semigroups, to prove the theorem via an application of our existence result, Theorem 2.1.

Let $\chi = \{\hat{x} = \{x_n\}_{n=0}^{\infty}: x_n \in X_n,\ n = 0, 1, 2, \cdots,$ and $\lim_{n\to\infty} x_n = x_0\}$ and for $\hat{x} \in \chi$ set $\|\hat{x}\| = \sup_n |x_n|_0$. Then $\|\cdot\|$ defines a norm on the linear vector space χ, and the space χ together with the norm $\|\cdot\|$ is a Banach space. Define the operator $A: \chi \to 2^{\chi}$ by

$$\begin{aligned}\mathrm{dom}(A) = \{\hat{x} = \{x_n\}_{n=0}^{\infty} \in \chi: {} & x_n \in \mathrm{Dom}(A_n), \text{ and for each } n = 1, 2, \cdots \text{ there} \\ & \text{exists a } y_n \in A_n x_n \text{ such that } \lim_{n\to\infty} y_n = y_0 \in A_0 x_0\},\end{aligned}$$

for $\hat{x} \in \mathrm{Dom}(A)$, $\hat{y} = \{y_n\}_{n=0}^{\infty} \in A\hat{x}$ if and only if $y_n \in A_n x_n$, $n = 0, 1, \cdots$ and $\lim_{n\to\infty} y_n = y_0$.

Define an essentially χ-valued function f on the interval $[0, T]$ by $f(t) = \{f_n(t)\}_{n=0}^{\infty}$. The assumptions on the f_n are such that $f_n(t) \to f_0(t)$ for almost every $t \in [0, T]$. However, by appropriately redefining on a set of measure zero, we may infer from the assumptions on the functions f_n that $f: [0, T] \to \chi$ with $f \in L_1(0, T; \chi)$.

It is readily seen that the operator $A + \omega_0 I$ is m-accretive. Let $\hat{x}^1 = \{x_n^1\}_{n=0}^{\infty}$, $\hat{x}^2 = \{x_n^2\}_{n=0}^{\infty} \in \mathrm{Dom}(A)$, and let $\hat{y}^1 = \{y_n^1\}_{n=0}^{\infty} \in A\hat{x}^1$ and $\hat{y}^2 = \{y_n^2\}_{n=0}^{\infty} \in A\hat{x}^2$. Since for each $n = 0, 1, 2, \cdots$, $A_n + \omega_0 I$ is assumed to be m-accretive, for $\lambda > 0$, we have

$$\begin{aligned}\|\hat{x}^1 - \hat{x}^2\| &= \sup_n |x_n^1 - x_n^2|_0 \leqq \sup_n |x_n^1 - x_n^2 + \lambda(y_n^1 + \omega_0 x_n^1 - (y_n^2 + \omega_0 x_n^2))|_0 \\ &= \|\hat{x}^1 - \hat{x}^2 + \lambda(\hat{y}^1 + \omega_0 \hat{x}^1 - (\hat{y}^2 + \omega_0 \hat{x}^2))\|,\end{aligned}$$

and therefore that $A + \omega_0 I$ is accretive. Now let $\hat{y} = \{y_n\}_{n=0}^{\infty} \in \chi$ and set $\hat{x} = \{x_n\}_{n=0}^{\infty}$ with $x_n = J(\lambda_0; A_n + \omega_0 I)y_n$, $n = 0, 1, 2, \cdots$ where λ_0 is chosen as in (2.5). It is immediately clear that for each $n = 0, 1, 2, \cdots$, $x_n \in \mathrm{Dom}(A_n) \subset X_n$. Since $\hat{y} \in \chi$ we have $\lim_{n\to\infty} y_n = y_0$ and therefore, by assumption (2.5), that $\lim_{n\to\infty} x_n = x_0$ or $\hat{x} \in \chi$. Setting $z_n = (y_n - (1 + \lambda_0\omega_0)x_n)/\lambda_0$, $n = 0, 1, 2, \cdots$, it follows that $z_n \in A_n x_n$ and $\lim_{n\to\infty} z_n = z_0 \in A_0 x_0$. We conclude that $\hat{x} \in \mathrm{Dom}(A)$, $(I + \lambda_0(A + \omega_0 I))\hat{x} \ni \hat{y}$, and that $\mathcal{R}(I + \lambda_0(A + \omega_0 I)) = \chi$.

We have shown that the operator A and the function f satisfy conditions (1) and (2) given in the statement of Theorem 2.1. Therefore, a unique nonlinear evolution system $\{U(t, s): 0 \leqq s \leqq t \leqq T\}$ on $\overline{\mathrm{Dom}(A)}$ corresponding to A and f can be constructed with $U(t, s) = \{U_n(t, s)\}_{n=0}^{\infty}$. Using assumption (2.5) it can be shown that $\overline{\mathrm{Dom}(A)} = \{\hat{x} = \{x_n\}_{n=0}^{\infty} \in \chi: x_n \in \bar{D}_n,\ n = 0, 1, 2, \cdots$ and $\lim_{n\to\infty} x_n = x_0\}$. Since $\mathcal{R}(U(t, s)) \subset \chi$, it

follows that

$$\lim_{n\to\infty} U_n(t,s)\phi_n = U_0(t,s)\phi_0, \qquad 0 \leqq s \leqq t \leqq T \tag{2.7}$$

whenever $\phi_n \in \bar{D}_n$ and $\lim_{n\to\infty} \phi_n = \phi_0 \in \bar{D}_0$. Since each of the operators A_n and the functions f_n satisfy conditions (1) and (2) of Theorem 2.1, unique nonlinear evolution systems $\{U_n(t,s): 0 \leqq s \leqq t \leqq T\}$ on $\bar{D}_n$ corresponding to A_n and f_n can be constructed. Recalling that $\overline{\mathrm{Dom}(A)} \subset \bigtimes_{n=0}^{\infty} \bar{D}_n$, we may define the family of operators $\{V(t,s): 0 \leqq s \leqq t \leqq T\}$ on $\overline{\mathrm{Dom}(A)}$ by

$$V(t,s)\hat{x} = \{V_n(t,s)x_n\}_{n=0}^{\infty} \equiv \{U_n(t,s)x_n\}_{n=0}^{\infty} \tag{2.8}$$

for $\hat{x} = \{x_n\}_{n=0}^{\infty} \in \overline{\mathrm{Dom}(A)}$. Uniqueness (see [9]) dictates that for each $n = 0, 1, 2, \cdots$, $U_n(t,s)x_n = V_n(t,s)x_n$ whenever $\{x_n\}_{n=0}^{\infty} \in \overline{\mathrm{Dom}(A)}$. This together with (2.7) and (2.8) establish (2.6). The fact that the convergence in (2.6) is uniform in t for $t \in [s, T]$ is argued exactly as it was for the convergence of approximations to nonlinear semigroups in the proof of Theorem 3.2 in [10].

We note that (2.5) is also a necessary condition for the conclusion to hold (see, for example, Theorem 1 of [14]).

3. An approximation theory for identification problems. Let H be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $|\cdot|$. Let V be a reflexive real Banach space with norm $\|\cdot\|$ and let V^* be its dual. (Our entire theory can be developed in complex spaces if necessary; see [6].) We denote the usual dual norm on V^* by $\|\cdot\|_*$ and assume that V is densely and continuously embedded in H with $|v| \leqq \mu\|v\|$, $v \in V$, for some positive constant μ. Identifying H with its dual, we obtain $V \subset H = H^* \subset V^*$. For $\phi \in V^*$ and $v \in V$ the duality pairing between ϕ and v is denoted by $\langle \phi, v \rangle$. When $\phi \in H$, its pairing with $v \in V$ agrees with the inner product of ϕ with v. It follows for $u \in H$ and $v \in V$ that $\|u\|_* \leqq \mu|u|$ and $\|v\|_* \leqq \mu^2\|v\|$. Let $\mathcal{Q}$ and Z be metric spaces and let Q be a nonempty, compact, subset of $\mathcal{Q}$. The spaces $\mathcal{Q}$ and Z, and the set Q are referred to as the parameter space, the observation space, and the admissible parameter set, respectively.

We recall that a single-valued operator $A: V \to V^*$ is hemicontinuous if $\lim_{t\to 0} A(u+tv) = Au$ for all $u, v \in V$ where the limit is taken in the weak sense.

For each $q \in Q$ let $A(q): V \to V^*$ be a single-valued, hemicontinuous (in general, nonlinear) operator satisfying:

(A) (Continuity): For each $v \in V$, the map $q \to A(q)v$ is continuous from $Q \subset \mathcal{Q}$ into V^*.

(B) (Equi V-monotonicity): There exist an $\omega \in R$ and an $\alpha > 0$, both independent of $q \in Q$, such that

$$\langle A(q)u - A(q)v, u - v \rangle + \omega|u - v|^2 \geqq \alpha\|u - v\|^2,$$

for every $u, v \in V$.

(C) (Equiboundedness): The operator $A(q)$ maps V bounded sets into V^* bounded sets, uniformly in $q \in \mathcal{Q}$. That is, if S is V bounded, then there exists M_S depending on S such that

$$\sup\{\|A(q)v\|_*: v \in S, q \in \mathcal{Q}\} \leqq M_S.$$

For each $q \in Q$, let $f(\cdot; q) \in L_1(0, T; H)$ and $u^0(q) \in H$, and assume that the mapping $q \to u^0(q)$ is continuous from $Q \subset \mathcal{Q}$ into H and that the mapping $q \to f(t; q)$ is continuous from $Q \subset \mathcal{Q}$ into H for almost every $t \in [0, T]$. Also, for every $z \in Z$, let $u \to \Phi(u; z)$ be a continuous map from $C(0, T; H)$ into R^+.

We consider parameter identification or inverse problems of the following form:

(ID) Given observations $z \in Z$, determine parameters $\bar{q} \in Q$ that minimize

$$\phi(q) = \Phi(u_0(q); z)$$

where $u_0(q) = u_0(\cdot\,; q)$ is a mild solution to the initial value problem

$$\dot{u}(t) + A(q)u(t) = f(t; q), \qquad 0 < t \leqq T, \tag{3.1}$$

$$u(0) = u^0(q) \tag{3.2}$$

corresponding to $q \in Q$.

By a mild solution to (3.1), (3.2) we mean a solution in the sense of Theorem 2.1. To be more precise, for each $q \in Q$ we define the operator $A_0(q)$: Dom $(A_p(q)) \subset H \to H$ to be the restriction of the operator $A(q)$ to the subset of V given by Dom $(A_0(q)) = \{v \in V\colon A(q) \in H\}$, and prove the following theorem.

THEOREM 3.1. *For each $q \in Q$ the operator $A_0(q)$: Dom $(A_0(q)) \subset H \to H$ is densely defined and the operator $A_0(q) + \omega I$ is m-accretive.*

Proof. We first show that for each $q \in Q$ the operator $A(q) + \omega I : V \to V^*$ is coercive. If $\{v_n\} \subset V$ with $\lim_{n\to\infty} \|v_n\| = \infty$, then from assumptions (B) and (C) we obtain

$$\begin{aligned}
&\lim_{n\to\infty} \langle (A(q) + \omega I)v_n, v_n\rangle / \|v_n\| \\
&\quad = \lim_{n\to\infty} \{(\langle A(q)v_n - A(q)\theta, v_n\rangle + \omega |v_n|^2)/\|v_n\| + \langle A(q)\theta, v_n\rangle/\|v_n\|\} \\
&\quad \geqq \lim_{n\to\infty} \{\alpha \|v_n\|^2/\|v_n\| - |\langle A(q)\theta, v_n\rangle|/\|v_n\|\} \\
&\quad \geqq \lim_{n\to\infty} \{\alpha \|v_n\| - \|A(q)\theta\|_*\} \geqq \lim_{n\to\infty} \alpha \|v_n\| - \beta = \infty
\end{aligned}$$

where θ denotes the zero vector in V and $\beta > \sup_{q\in Q} \|A(q)\theta\|_*$. It follows that for each $\lambda > 0$, the operator $I + \lambda(A(q) + \omega I)\colon V \to V^*$ is monotone, everywhere defined on V, hemicontinuous, and coercive. Consequently, $\mathcal{R}(I + \lambda(A(q) + \omega I)) = V^*$ (see Barbu [6, Thm. II.1.3]) and therefore $\mathcal{R}(I + \lambda(A_0(q) + \omega I)) = H$. Also, for $u, v \in$ Dom $(A_0(q))$, we may use assumption (B) to conclude that

$$\begin{aligned}
\left(1 + \frac{\lambda\alpha}{\mu^2}\right)|u - v|^2 &\leqq |u - v|^2 + \lambda\alpha \|u - v\|^2 \\
&\leqq |u - v|^2 + \lambda\langle (A(q) + \omega I)u - (A(q) + \omega I)v, u - v\rangle \\
&= \langle (I + \lambda(A(q) + \omega I))u - (I + \lambda(A(q) + \omega I))v, u - v\rangle \\
&\leqq |(I + \lambda(A_0(q) + \omega I))u - (I + \lambda(A_0(q) + \omega I))v|\,|u - v|
\end{aligned}$$

or

$$|u - v| \leqq |u - v + \lambda((A_0(q) + \omega I)u - (A_0(q) + \omega I)v)|,$$

which proves that $A_0(q) + \omega I$ is m-accretive on Dom $(A_0(q)) \subset H$.

In light of Theorem 3.1, we may apply Theorem 2.1 with $X = H$, $A = A_0(q)$ and $f = f(\cdot, q)$. We conclude that there exists a unique nonlinear evolution system $\{U_0(t, s; q)\colon 0 \leqq s \leqq t \leqq T\}$ on $\overline{\text{Dom}\,(A_p(q))} \subset H$ satisfying (i)-(iii) of Theorem 2.1. The mild solution $u_0(\cdot\,; q)\colon [0, T] \to H$ to the initial value problem (3.1), (3.2) is given by $u_0(t; q) = U_0(t, 0; q)u^0(q)$ for $t \in [0, T]$.

While it is not true in general that $\overline{\text{Dom}\,(A_0(q))} = H$, we can, in several special cases of importance, argue this equality. We present two such cases.

LEMMA 3.1. *If the operator $A(q)$ satisfies, in addition to* (A), (B), (C), *the following condition, then* $\overline{\text{Dom}(A_0(q))} = H$: *$A(q)$ takes H bounded sets into V^* bounded sets.*

Proof. To show $\overline{\text{Dom}(A_0(q))} = H$, we let $u \in H$ and for each $n = 1, 2, \cdots$, we set $u_n = J(1/n; A_0(q) + \omega I)u \in \text{Dom}(A_0(q))$. Then, arguing as we have above, we find that

$$|u_n|^2 + (1/n)\alpha\|u_n\|^2 \leqq \langle u - (1/n)A(q)\theta, u_n\rangle$$
$$\leqq |u||u_n| + (1/n)\|A(q)\theta\|_*\|u_n\|$$

where θ is again the zero vector in V. But then

$$\left(\frac{1}{2}\right)|u_n|^2 + \left(\frac{1}{n}\right)\left(\frac{\alpha}{2}\right)\|u_n\|^2 \leqq \left(\frac{1}{2}\right)|u|^2 + \left(\frac{1}{n}\right)\left(\frac{1}{2\alpha}\right)\|A(q)\theta\|_*^2 \tag{3.3}$$
$$\leqq \left(\frac{1}{2}\right)|u|^2 + \left(\frac{1}{n}\right)\left(\frac{\beta^2}{2\alpha}\right),$$

from which it immediately follows that the u_n are uniformly bounded in H. Indeed, from (3.3) we see that $(1/n)\|u_n\|^2$, and hence $\|u_n\|/\sqrt{n}$ is bounded so that $\|u_n\|/n \to 0$ as $n \to \infty$.

Thus we find that (using the fact that $\{A_0(q)u_n\}$ is V^* bounded)

$$\|u_n - u\|_* = \left(\frac{1}{n}\right)\|(A_0(q) + \omega I)u_n\|_*$$
$$\leqq \frac{1}{n}\{\|(A_0(q)u_n\|_* + \mu^2\omega\|u_n\|\} \tag{3.3a}$$
$$\leqq \frac{1}{n}\{K + \mu^2\omega\|u_n\|\}.$$

Since the last term in the estimate above tends to zero as $n \to \infty$, we find $u_n \to u$ in V^* as $n \to \infty$. This, together with the fact that V is dense in H imply that $u_n \to u$ weakly in H as $n \to \infty$. Since $A_0(q) + \omega I$ is m-accretive, $\overline{\text{Dom } A_0(q)}$ is convex from which $\overline{\text{Dom}(A_0(q))} = H$ immediately follows.

We note that the additional condition ($A(q)$ takes H bounded sets into V^* bounded sets) of Lemma 3.1 allows for a number of nonlinearities that arise in applications. For example, in a typical case ($H = L_2(0, 1)$, $V = H^2(0, 1)$) this condition is satisfied by A given by $(Av)(x) = |v(x)|^\alpha \operatorname{sgn} v(x)$ for $0 \leqq \alpha \leqq 2$.

We can also establish the results of Lemma 3.1 by strengthening condition (C).

LEMMA 3.2. *If the operator $A(q)$ satisfies* (A), (B), *and the following condition, then* $\overline{\text{Dom}(A_0(q))} = H$:

($\tilde{\text{C}}$) *There exists a constant $\beta > 0$, independent of $q \in Q$ such that $\|A(q)v\|_* \leqq \beta\{\|v\| + 1\}$ for every $v \in V$.*

Proof. The arguments are exactly the same as those for Lemma 3.1 except that (3.3a) must be replaced by

$$\|u_n - u\|_* \leqq \frac{1}{n}\|(A_0(q) + \omega I)u_n\|_* \leqq \frac{1}{n}\{(\beta + \omega\mu^*)\|u_n\| + \beta\}.$$

Remark. Under additional hypotheses on $f(\cdot; q)$ and $u^0(q)$ other existence results can be applied to obtain somewhat different notions of a solution to the initial value problem (3.1), (3.2). For example (see [6, pp. 140–144]), if $f(\cdot; q) \in W^{1,1}(0, T; H)$ and $u^0(q) \in \text{Dom}(A_0(q))$, then there exists a unique $u(\cdot, q): [0, T] \to V$ satisfying $u(\cdot; q) \in W^{1,\infty}(0, T; H)$, $A(q)u(\cdot; q) \in L_\infty(0, T; H)$ and $\dot{u}(t; q) + A(q)u(t; q) = f(t; q)$ for

almost every $t \in [0, T]$. Or, if $u^0(q) \in H$ and $f(\cdot; q) \in L_2(0, T; V^*)$ then there exists a unique $u(\cdot; q)$ that is V^*-valued absolutely continuous almost everywhere on $[0, T]$, $u(\cdot; q) \in C(0, T; H) \cap L_2(0, T; V)$, $\dot{u}(\cdot; q) \in L_2(0, T; V^*)$ and $\dot{u}(t; q) + A(q)u(t; q) = f(t; q)$, for almost every $t \in [0, T]$. If, in addition, the mapping $t \to t^\gamma f'(t; q)$ is an element in $L_2(0, T; V^*)$ for some $\gamma \geqq 1$, then the mapping $t \to t^\gamma \dot{u}(t; q)$ is in $L_2(0, T; V) \cap L_\infty(0, T; H)$. In particular, when $f(\cdot; q) = 0$, the nonlinear semigroup $\{S_0(t; q)\colon 0 \leqq t \leqq T\}$ on H defined by $S_0(t; q) = U_0(t; 0; q)$, $t \in [0, T]$, with generator $-A_0(q)$ behaves as does a holomorphic linear semigroup in that it smooths. That is, $S_0(t; q)u^0(q) \in \text{Dom}\,(A_0(q))$, $t \in (0, T]$, and the mapping $t \to t(d/dt)S(t; q)u^0(q)$ is an element in $L_\infty(0, T; H)$ for every $u^0(q) \in H$. Also, some generalizations are possible. For example, in assumption (B), the term $\alpha\|u - v\|^2$ can be replaced by a term of the form $\alpha(\|u - v\|)\|u - v\|$ where $\alpha(\cdot)$ is a continuous, strictly increasing function on $[0, \infty)$ satisfying $\alpha(0) = 0$ and $\lim_{x \to \infty} \alpha(x) = \infty$. Or, the terms $\|u - v\|^2$ in (B) and $\|v\|$ in ($\tilde{C}$) can be replaced by $\|u - v\|^p$ and $\|v\|^{p-1}$, respectively, for any $p \geqq 2$.

The development of computational methods for the solution of the infinite-dimensional optimization problem (ID) requires the finite-dimensional approximation of the abstract initial value problem (3.1), (3.2). The general framework that we are proposing is based on a classical Galerkin approach. For each $n = 1, 2, \cdots$ let H_n denote a finite-dimensional subspace of H that is a subset of V. Let $P_n : H \to H_n$ denote the orthogonal projection of H onto H_n with respect to the $\langle \cdot, \cdot \rangle$ inner product. We assume that the approximating subspaces H_n, and the projections P_n satisfy

(D) For each $v \in V$, $\lim_{n \to \infty} \|P_n v - v\| = 0$.

Note that assumption (D) and V densely and continuously embedded in H imply that $\lim_{n \to \infty} |P_n u - u| = 0$ for each $u \in H$.

For each $q \in Q$ and $n = 1, 2, \cdots$ we define the single-valued operator $A_n(q)\colon H_n \to H_n$ by $A_n(q)u_n = v_n$ for $u_n \in H_n$ where v_n satisfies

$$\langle A(q)u_n, w_n \rangle = \langle v_n, w_n \rangle, \qquad w_n \in H_n.$$

That $A_n(q)$ is a well-defined operator from H_n into H_n follows from the Riesz Representation Theorem applied to the Hilbert space H_n and the bounded linear functional $\langle A(q)u_n, \cdot \rangle$ on H_n. Also, define $f_n(\cdot; q)\colon [0, T] \to H_n$ and $u_n^0(q) \in H_n$ by $f_n(t; q) = P_n f(t; q)$, $0 \leqq t \leqq T$, and $u_n^0(q) = P_n u^0(q)$, respectively. Note that $f_n(\cdot; q) \in L_1(0, T; H_n) \subset L_1(0, T; H)$ and that $|f_n(t; q)| \leqq |f(t; q)|$ for $q \in Q$ and almost every $t \in [0, T]$.

We consider the sequence of approximating identification problems given by the following:

(ID_n) Given observations $z \in Z$, determine parameters $\bar{q}_n \in Q$ that minimize

$$\phi_n(q) = \Phi(u_n(q); z)$$

where $u_n(q) = u_n(\cdot; q)$ is a mild solution to the initial value problem in H_n

$$\dot{u}_n(t) + A_n(q)u_n(t) = f_n(t; q), \qquad 0 < t \leqq T, \tag{3.4}$$

$$u_n(0) = u_n^0(q) \tag{3.5}$$

corresponding to $q \in Q$.

From the definition of the $A_n(q)$ and the assumptions (B) and (C) on $A(q)$, using arguments analogous to those used to prove Theorem 3.1, it can be shown that the operators $A_n(q) + \omega I$ are m-accretive on H_n. It then follows from Theorem 2.1 that

for each $n = 1, 2, \cdots$ there exists a unique nonlinear evolution system $\{U_n(t, s; q): 0 \leqq s \leqq t \leqq T\}$ on H_n satisfying (i)-(iii) in the statement of that theorem with $X = H_n$, $f(t) = f_n(t; q)$, and $x^0 = u_n^0(q)$. The mild solution to the initial value problem (3.4), (3.5) is given by $u_n(t; q) = U_n(t, 0; q)u_n^0(q)$, $t \in [0, T]$.

If we assume for the moment that the approximating identification problems (ID_n) have solutions $\bar{q}_n \in Q$, then it is desirable that they in some sense approximate a solution $\bar{q}$ to the original identification problem (ID). This is in fact the case. For suppose that it can be shown that for any sequence $\{q_n\} \subset Q$ with $\lim_{n\to\infty} q_n = q \in Q$, we have

$$\lim_{n\to\infty} u_n(q_n) = u_0(q_0) \quad \text{in } C(0, T; H). \tag{3.6}$$

Then $\{\bar{q}_n\} \subset Q$ and Q a compact subset of the metric space $\mathcal{Q}$ imply that there exist a subsequence $\{\bar{q}_{n_j}\} \subset \{\bar{q}_n\}$ and a $\bar{q} \in Q$ such that $\lim_{j\to\infty} \bar{q}_{n_j} = \bar{q}$. For any $q \in Q$ the continuity of Φ implies

$$\begin{aligned}
\phi(\bar{q}) &= \Phi(u_0(\bar{q}); z) = \Phi\left(\lim_{j\to\infty} u_{n_j}(\bar{q}_{n_j}); z\right) \\
&= \lim_{j\to\infty} \Phi(u_{n_j}(\bar{q}_{n_j}); z) = \lim_{j\to\infty} \phi_{n_j}(\bar{q}_{n_j}) \\
&\leqq \lim_{j\to\infty} \phi_{n_j}(q) = \lim_{j\to\infty} \Phi(u_{n_j}(q); z) \\
&= \Phi\left(\lim_{j\to\infty} u_{n_j}(q); z\right) = \Phi(u_0(q); z) \\
&= \phi(q).
\end{aligned}$$

Note that in the discussion above we did not assume that a solution to problem (ID) exists. But rather we have shown that the existence of solutions $\bar{q}_n$ to the approximating problems (ID_n) and (3.6) imply the existence of a solution $\bar{q}$ to problem (ID). When the solution to problem (ID) is unique, the sequence $\{\bar{q}_n\}$ itself converges to $\bar{q}$.

The existence of a solution $\bar{q}_n$ to problem (ID_n) for each $n = 1, 2, \cdots$ will follow from the compactness of Q and the continuity of Φ once the following continuous dependence result has been established: $\lim_{m\to\infty} u_n(q_m) = u_n(q_0)$ in $C(0, T; H_n)$ whenever $\{q_m\} \subset Q$ with $\lim_{m\to\infty} q_m = q_0$. Although continuous dependence for the finite-dimensional systems (3.4), (3.5) could be demonstrated via a modification to any one of a number of familiar continuous dependence results for ordinary differential equations (see, for example, Hale [12, Thm. I.3.4]), it is also easily handled with the approximation theory developed in the previous section. This and the convergence in (3.6) are addressed in the following theorem.

THEOREM 3.2. *If assumptions* (A)-(D) *hold, then*

(a) *If* $\{q_n\} \subset Q$ *with* $\lim_{n\to\infty} q_n = q_0$ *then* $\lim_{n\to\infty} u_n(q_n) = u_0(q_0)$ *in* $C(0, T; H)$; *and*

(b) *If* $\{q_m\} \subset Q$ *with* $\lim_{m\to\infty} q_m = q_0$ *then* $\lim_{m\to\infty} u_n(q_m) = u_n(q_0)$ *for each* $n \in Z^+$.

Proof. Assumption (D) and the continuity of the map $q \to u^0(q)$ from Q into H imply $\lim_{n\to\infty} u_n^0(q_n) = u^0(q_0)$ in H. Hence, we will have verified (a) if we can show that $\lim_{n\to\infty} U_n(t, s; q_n)w_n = U_0(t, s; q_0)w_0$, $0 \leqq s \leqq t \leqq T$, uniformly in t for $t \in [s, T]$ whenever $w_n \in H_n$ with $\lim_{n\to\infty} w_n = w_0 \in H$. We argue this using Theorem 2.2. Note that assumption (D) implies $\lim_{n\to\infty} H_n \supset H$ and assumption (D) together with the assumed continuity of the map $q \to f(t; q)$ from $Q \subset \mathcal{Q}$ into H for almost every $t \in [0, T]$ imply $\lim_{n\to\infty} f_n(t; q_n) = f(t; q_0)$ in H for almost every $t \in [0, T]$ with the $f_n(\cdot; q_n)$

dominated by a function $g \in L_1(0, T; H)$, which is independent of n. Thus, we need only to demonstrate that for some $\lambda_0 > 0$, we have

$$\lim_{n\to\infty} J(\lambda_0; A_n(q_n)+\omega I)w_n = J(\lambda_0; A_0(q_0)+\omega I)w_0 \tag{3.7}$$

in H whenever $w_n \in H_n$, $n \in Z^+$ with $\lim_{n\to\infty} w_n = w_0$.

Let $\lambda_0 > 0$ and set $v_n = J(\lambda_0; A_n(q_n)+\omega I)w_n$ and $v_0 = J(\lambda_0; A_0(q)+\omega I)w_0$. We first show that $\|v_n\|$ is uniformly bounded in n. From assumption (B) we obtain

$$\begin{aligned}
\lambda_0\alpha\|v_n\|^2 &\leqq \lambda_0\omega|v_n|^2 + \lambda_0\langle A(q_n)v_n - A(q_n)\theta, v_n\rangle \\
&= \langle (I+\lambda_0(A_n(q_n)+\omega I))v_n, v_n\rangle - |v_n|^2 \\
&\quad + \lambda_0\langle A(q_0)\theta - A(q_n)\theta, v_n\rangle - \lambda_0\langle A(q_0)\theta, v_n\rangle \\
&= \langle w_n, v_n\rangle - |v_n|^2 + \lambda_0\langle A(q_0)\theta - A(q_n)\theta, v_n\rangle - \lambda_0\langle A(q_0)\theta, v_n\rangle \\
&\leqq \|w_n\|_*\|v_n\| + \lambda_0\|A(q_0)\theta - A(q_n)\theta\|_*\|v_n\| + \lambda_0\|A(q_0)\theta\|_*\|v_n\|
\end{aligned}$$

where θ denotes the zero vector in V. This estimate together with assumption (C) yields

$$\|v_n\| \leqq (\lambda_0\alpha)^{-1}\mu|w_n| + \alpha^{-1}\|A(q_n)\theta - A(q_0)\theta\|_* + \alpha^{-1}M_{\{\theta\}}.$$

Recalling assumption (A) and that $\lim_{n\to\infty} w_n = w_0$ in H, we find that the desired uniform bound on $\|v_n\|$ has been established.

Once again, from assumption (B), we find

$$\begin{aligned}
\lambda_0\alpha\|v_n - v_0\|^2 &\leqq \lambda_0\omega|v_n - v_0|^2 + \lambda_0\langle A(q_n)v_n - A(q_n)v_0, v_n - v_0\rangle \\
&= \lambda_0\omega|v_n - v_0|^2 + \lambda_0\langle A(q_n)v_n - A(q_0)v_0, v_n - P_nv_0\rangle \\
&\quad + \lambda_0\langle A(q_n)v_n - A(q_0)v_0, P_nv_0 - v_0\rangle \\
&\quad + \lambda_0\langle A(q_0)v_0 - A(q_n)v_0, v_n - v_0\rangle \\
&= \lambda_0\omega\langle P_nv_0 - v_0, v_n - v_0\rangle + \langle (I+\lambda_0(A_n(q_n)+\omega I))v_n \\
&\quad - (I+\lambda_0(A_0(q_0)+\omega I))v_0, v_n - P_nv_0\rangle + \langle v_0 - v_n, v_n - P_nv_0\rangle \\
&\quad + \lambda_0\langle A(q_n)v_n - A(q_0)v_0, P_nv_0 - v_0\rangle \\
&\quad + \lambda_0\langle A(q_0)v_0 - A(q_n)v_0, v_n - v_0\rangle \\
&= \lambda_0\omega\langle P_nv_0 - v_0, v_n - v_0\rangle + \langle w_n - w_0, v_n - P_nv_0\rangle - |v_n - P_nv_0|^2 \\
&\quad + \lambda_0\langle A(q_n)v_n - A(q_0)v_0, P_nv_0 - v_0\rangle \\
&\quad + \lambda_0\langle A(q_0)v_0 - A(q_n)v_0, v_n - v_0\rangle \\
&\leqq \lambda_0\omega\|P_nv_0 - v_0\|_*\|v_n - v_0\| + \|w_n - w_0\|_*\|v_n - v_0\| \\
&\quad + \|w_n - w_0\|_*\|P_nv_0 - v_0\| + \lambda_0\|A(q_n)v_n - A(q_0)v_0\|_*\|P_nv_0 - v_0\| \\
&\quad + \lambda_0\|A(q_0)v_0 - A(q_n)v_0\|_*\|v_n - v_0\|.
\end{aligned}$$

The estimate $ab \leqq (1/2\eta)a^2 + (\eta/2)b^2$ for any $\eta > 0$ and assumption (C) allow us to argue that (here we let M_S be the bound of (C) for $S = \{\|v_n\|, \|V_0\|\}$)

$$\begin{aligned}
\frac{\lambda_0\alpha}{2}\|v_n - v_0\|^2 &\leqq \frac{3\omega^2\lambda_0}{2\omega}\|P_nv_0 - v_0\|_*^2 + \frac{3}{2\lambda_0\alpha}\|w_n - w_0\|_*^2 + \|w_n - w_0\|_*\|P_nv_0 - v_0\| \\
&\quad + \lambda_0\|A(q_n)v_n - A(q_0)v_0\|_*\|P_nv_0 - v_0\| + \frac{3\lambda_0}{2\alpha}\|A(q_0)v_0 - A(q_n)v_0\|_*^2 \\
&\leqq \left\{\frac{3\omega^2\lambda_0\mu^4}{2\alpha} + \frac{(\lambda_0+1)}{2}\right\}\|P_nv_0 - v_0\|^2 + \left\{\frac{3\mu^2}{2\lambda_0\alpha} + \frac{\mu^2}{2}\right\}|w_n - w_0|^2 \\
&\quad + \lambda_0 2M_S\|P_nv_0 - v_0\| + \frac{3\lambda_0}{2\alpha}\|A(q_0)v_0 - A(q_n)v_0\|_*^2.
\end{aligned}$$

From this, $\lim_{n\to\infty} w_n = w_0$ in H, and assumptions (A) and (D), we can conclude, that $\lim_{n\to\infty} v_n = v_0$ in V and that (3.7) holds.

An analogous, but somewhat simpler argument can be used to verify (b). We use Theorem 2.2 to show that for $n \in Z^+$ fixed, $\lim_{m\to\infty} U_n(t, s; q_m)w_m = U_n(t, s; q_0)w_0$, $0 \leqq s \leqq t \leqq T$, uniformly in t for $t \in [s, T]$ whenever $w_m, w_0 \in H_n$ with $\lim_{m\to\infty} w_m = w_0$ in H. Clearly, $\lim_{m\to\infty} f_n(\cdot; q_m) = f_n(\cdot; q_0)$ in $L_1(0, T; H_n)$ so that we need only to show that for some $\lambda_0 > 0$,

$$\lim_{m\to\infty} J(\lambda_0; A_n(q_m) + \omega I)w_m = J(\lambda_0; A_n(q_0) + \omega I)w_0$$

in H whenever $\lim_{m\to\infty} w_m = w_0$ in H. Let $v_m = J(\lambda_0; A_n(q_m) + \omega I)w_m$ and $v_0 = J(\lambda_0; A_n(q_0) + \omega I)w_0$. Then from assumption (B)

$$\begin{aligned}
\lambda_0\alpha \|v_m - v_0\|^2 &\leqq \lambda_0\omega |v_m - v_0|^2 + \lambda_0\langle A(q_m)v_m - A(q_m)v_0, v_m - v_0\rangle \\
&= \langle (I + \lambda_0(A_n(q_m) + \omega I))v_m - (I + \lambda_0(A_n(q_0) + \omega I))v_0, v_m - v_0\rangle \\
&\quad - |v_m - v_0|^2 + \lambda_0\langle A(q_0)v_0 - A(q_m)v_0, v_m - v_0\rangle \\
&= \langle w_m - w_0, v_m - v_0\rangle - |v_m - v_0|^2 + \lambda_0\langle A(q_0)v_0 - A(q_m)v_0, v_m - v_0\rangle \\
&\leqq \|w_m - w_0\|_* \|v_m - v_0\| + \lambda_0 \|A(q_0)v_0 - A(q_m)v_0\|_* \|v_m - v_0\|
\end{aligned}$$

or

$$\|v_m - v_0\| \leqq \frac{\mu}{\lambda_0\alpha} |w_m - w_0| + \frac{1}{\alpha} \|A(q_0)v_0 - A(q_m)v_0\|_*.$$

Assumption (A) and $\lim_{m\to\infty} w_m = w_0$ in H yield the desired result and the theorem is proved.

Remark. In practice, the approximating identification problems (ID_n) are solved using standard iterative search techniques (for example, steepest descent, Newton's method, etc.) requiring the evaluation of $\phi_n(q)$ for $q \in Q$ at each step. This in turn requires the integration of the finite-dimensional initial value problem (3.4), (3.5). Once a basis for H_n has been chosen, the solution to (3.4), (3.5) can be computed using any standard numerical integrator for ordinary differential systems. Also, the parameter space $\mathcal{Q}$ and the admissible parameter set Q are frequently functional in nature and are infinite-dimensional. When this is the case, the set Q must also be discretized. Suppose that for each $m = 1, 2, \cdots$, $I^m : Q \subset \mathcal{Q} \to \mathcal{Q}$ is a continuous map with finite-dimensional range and that $\lim_{m\to\infty} I^m(q) = q$ with the convergence uniform in q for $q \in Q$. Set $Q^m = I^m(Q)$ (note that Q^m is a compact subset of $\mathcal{Q}$) and consider the identification problems (ID_n^m) defined to be the problems (ID_n) with Q replaced by Q^m. It is clear that each of these problems admits a solution $\bar{q}_n^m$ and it is not difficult to argue that there exists a subsequence $\{\bar{q}_{n_k}^{m_j}\} \subset \{\bar{q}_n^m\}$ with $\lim_{j,k\to\infty} \bar{q}_{n_k}^{m_j} = \bar{q}$, $\bar{q}$ a solution to problem (ID) (see, for example, [4]). Once bases for H_n and the range of I^m have been chosen, problem (ID_n^m) involves the minimization of a functional over a compact subset of Euclidean space subject to finite-dimensional constraints.

Remark (Nonautonomous systems). Theorems 2.1 and 2.2 remain valid for certain classes of temporally inhomogeneous or time-dependent operators $A = A(t)$. To be more precise, the family of operators $A(t) : X \to 2^X$ must be m-accretive on X for almost every $t \in [0, T]$ and must satisfy

$$|J(\lambda; A(t))x - J(\lambda; A(s))x|_X \leqq \lambda |h(t) - h(s)|_X L(|x|_X) \tag{3.8}$$

for each $x \in X$, every λ satisfying $0 < \lambda \leqq \lambda_0$ for some $\lambda_0 > 0$, some $h \in L_1(0, T; X)$, some continuous, nondecreasing function $L: [0, \infty) \to [0, \infty)$ and almost every $t, s \in [0, T]$ (see [8], [9]). (Note that for simplicity we have taken $\omega = 0$; however, the discussion to follow remains valid for any $\omega \in R$.) The primary motivation for developing the framework outlined above was to define readily verifiable conditions on the operators $A(q): V \to V^*$ that, if satisfied, would (i) also automatically be satisfied by the Galerkin approximation $A_n(q)$ and (ii) lead to the desired convergence of solutions to the approximating identification problems to a solution to problem (ID). The natural assumption to add to (A)-(C) that certainly satisfies criterion (i) and that could conceivably lead to an estimate of the form (3.8) in H is that

$$\|A(t; q)v - A(s; q)v\|_* \leqq |h(t) - h(s)|\tilde{L}(|v|) \tag{3.9}$$

for each $v \in V$, almost every $s, t \in [0, T]$ and some $h \in L_1(0, T; H)$ and some continuous nondecreasing $\tilde{L}: [0, \infty) \to [0, \infty)$, both of which do not depend on $q \in Q$. Unfortunately, however, we can only show that (3.9) leads to an estimate of the form

$$|J(\lambda; A_0(t; q))u - J(\lambda; A_0(s; q))u| \leqq \sqrt{\lambda}\, |h(t) - h(s)| L(|u|) \tag{3.10}$$

for each $u \in H$. Moreover, it is not clear to us how, or if, the proof of the fundamental Theorem 2.1 given in [9] could be modified so that (3.10) would suffice. We have explored alternative approaches and developed other techniques for treating the nonautonomous case (for example, in the linear case, based on some ideas in Tanabe [18], and in the strongly monotone case, via a variational formulation which can be found in Barbu [6]). These results will appear soon in forthcoming papers.

4. Applications and examples. We briefly describe some classes of systems to which the general framework developed in the previous section applies. In our discussion below we consider theoretical aspects only. Implementation questions will be treated and the results of our numerical studies will be reported on elsewhere.

Example 4.1 (Linear regularly dissipative operators). The approximation theory for inverse problems for systems involving linear regularly dissipative operators was treated in detail by Banks and Ito in, and is the central focus of, [2] and [3]. We show here that the linear theory is a special case of the nonlinear theory given in § 3.

Let the spaces H, V, V^*, and $\mathcal{Q}$ and the set Q be as they have been defined above. For each $q \in Q$ let $a(q)(\cdot, \cdot)$ be a sesquilinear form defined on $V \times V$, which satisfies the following conditions:

(A′) For each $v \in V$ the mapping $q \to a(q)(\cdot, v)$ is continuous from $Q \subset \mathcal{Q}$ into V^*. That is, given $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\sup_{\substack{u \in V \\ \|u\| = 1}} |a(q_0)(u, v) - a(q)(u, v)| < \varepsilon$$

whenever $d(q_0, q) < \delta$ where d denotes the metric on $\mathcal{Q}$.

(B′) There exist an $\omega \in R$ and an $\alpha > 0$, both independent of $q \in Q$, for which $a(q)(v, v) + \omega |v|^2 \geqq \alpha \|v\|^2$ for every $v \in V$.

(C′) There exists a constant $\beta > 0$, independent of $q \in Q$, such that

$$|a(q)(u, v)| \leqq \beta \|u\| \, \|v\| \quad \text{for every } u, v \in V.$$

When conditions (A′)-(C′) are satisfied it is not difficult to argue that for each $q \in Q$ an operator $A(q) \in \mathcal{L}(V, V^*)$ can be defined by

$$[A(q)v](u) = \langle A(q)v, u \rangle = a(q)(u, v),$$

$u, v \in V$ and that $A(q): V \to V^*$ satisfies (A)-(C). It then follows from Theorems 2.1 and 3.1 that there exists a unique nonlinear evolution system $\{U_0(t, s; q): 0 \leqq s \leqq t \leqq T\}$ on H corresponding to the initial value problem

$$\dot{u}(t) + A_0(q)u(t) = f(t; q), \qquad 0 < t \leqq T,$$
$$u(0) = u^0(q)$$

where for each $q \in Q$, $f(\cdot; q) \in L_1(0, T; H)$, $u^0(q) \in H$, and $A_0(q)$: Dom $(A_0(q)) \subset H \to H$ is the restriction of $A(q)$ to the set Dom $(A_0(q)) = \{v \in V: A(q)v \in H\}$. The operator $-A_0(q)$ is the infinitesimal generator of an analytic semigroup $\{T_0(t; q): t \geqq 0\}$ on H (see [18]) and for $\phi \in H$

$$U_0(t, s; q)\phi = T_0(t-s; q)\phi + \int_s^t T_0(t-\tau; q)f(\tau; q)\, d\tau. \tag{4.1}$$

It can be shown that the semigroup $\{T_0(t; q): t \geqq 0\}$ admits an extension $\{T(t; q): t \geqq 0\}$, which is an analytic semigroup on V^* with generator $A(q): V \subset V^* \to V^*$. Also the restriction of $\{T_0(t; q): t \geqq 0\}$ to V, call it $\{\tilde{T}(t; q): t \geqq 0\}$, is an analytic semigroup on V with generator $\tilde{A}(q)$: Dom $(\tilde{A}(q)) \subset V \to V$, the restriction of $A(q)$ to the set Dom $(\tilde{A}(q)) = \{v \in V: A(q)v \in V\}$ (see [3], [18]). Consequently, with appropriate assumptions on $f(\cdot; q)$, the evolution system $\{U_0(t, s; q): 0 \leqq s \leqq t \leqq T\}$ admits an extension $\{U(t, s; q): 0 \leqq s \leqq t \leqq T\}$, which is an evolution system on V^* and a restriction $\{\tilde{U}(t, s; q): 0 \leqq s \leqq t \leqq T\}$, which is an evolution system on V.

It is clear from (4.1) that when $A(q)$ is linear, we may take $f(\cdot; q) \equiv 0$ and consider only the approximation of the semigroup $\{T_0(t; q): t \geqq 0\}$. For each $n = 1, 2, \cdots$ let the finite-dimensional subspaces H_n of H and the corresponding orthogonal projections P_n be as they were defined in § 3 and assume that condition (D) is satisfied. Denote the Galerkin approximations to $A(q)$ (i.e., the restriction of $A(q)$ to an operator from H_n into $H_n^* = H_n$) by $A_n(q)$ and set $T_n(t; q) = \exp(-tA_n(q))$, $t \geqq 0$. Theorem 3.2 then implies that

$$\lim_{n\to\infty} |T_n(t; q_n)P_n u^0(q_n) - T_0(t; q_0)u^0(q_0)| = 0 \tag{4.2}$$

uniformly in t, for $t \in [0, T]$ whenever $\{q_n\} \subset Q$ with $\lim_{n\to\infty} q_n = q_0 \in Q$, and the mapping $q \to u^0(q)$ is continuous from $Q \subset \mathcal{Q}$ into H. In addition, recalling that we required that $H_n \subset V$ for all $n = 1, 2, \cdots$, an inspection of the proof of Theorem 3.2 reveals that in the linear case with the existence of the semigroup $\{\tilde{T}(t; q): t \geqq 0\}$ on V, we may apply Theorem 2.2 with $X = V$ and conclude that

$$\lim_{n\to\infty} \|T_n(t; q_n)P_n u^0(q_n) - \tilde{T}(t; q_0)u^0(q_0)\| = 0 \tag{4.3}$$

uniformly in t for $t \in [0, T]$ whenever $\lim_{n\to\infty} q_n = q_0$, $u^0(q) \in V$ and the map $q \to u^0(q)$ is continuous from Q into V (see also [3]). Then for $\phi \in H$, setting

$$U_n(t, s; q)P_n\phi = T_n(t-s; q)P_n\phi + \int_s^t T_n(t-\tau; q)P_n f(\tau; q)\, d\tau$$

under appropriate assumptions on $f(\cdot; q)$, (4.2) and (4.3) continue to hold with $T_n(t; q)$, $T_0(t; q)$, and $\tilde{T}(t; q)$ replaced by $U_n(t, s; q)$, $U_0(t, s; q)$, and $\tilde{U}(t, s; q)$, respectively, with the convergence being uniform in t, for $t \in [s, T]$. Hence the linear theory and results of [3] are a special case of the nonlinear theory of § 3.

We note that in the context of the identification problem, the fact that the stronger V-convergence given in (4.3) can be obtained is significant. Indeed, (4.3) permits the

relaxation of the continuity assumption on the performance index Φ to the requirement that for each $z \in Z$, the mapping $u \to \Phi(u, z)$ be continuous from $C([0, T]; V)$ into R^+. This can have the effect of significantly enlarging the class of allowable observations (e.g., see [1]). For example, in the case of a one-dimensional parabolic system formulated in $H = L_2$ with V in H^1, spatially discrete (i.e., pointwise, as opposed to distributed in space) measurements will suffice (see [3] and [5]).

Among the class of linear regularly dissipative operators that arise from a form satisfying (A')-(C') are the familiar elliptic partial differential operators on L_2. Briefly, let Ω be a region in R^l and let $\mathcal{Q} = \times_{n=1}^{l^2+l+1} L_\infty(\Omega)$. Let Q be a compact subset of $\mathcal{Q}$ with the property that if $q = \{(a_{ij}, b_i, c): i, j = 1, \cdots, l\} \in Q$, then for some $\alpha > 0$ independent of $q \in Q$,

$$\sum_{i,j=1}^{l} a_{ij}(x)\zeta_i\zeta_j \geqq \alpha|\zeta|^2$$

for every $x \in \Omega$, and every $\zeta \in R^l$. For $q \in Q$ and $u, v \in H^1(\Omega)$ set

$$a(q)(u, v) = \int_\Omega \left\{ \sum_{i,j=1}^{l} a_{ij}(x) \frac{\partial u(x)}{\partial x_i} \frac{\partial v(x)}{\partial x_j} + \sum_{i=1}^{l} b_i(x) \frac{\partial u(x)}{\partial x_i} v(x) + c(x)u(x)v(x) \right\} dx$$

with $H = L_2(\Omega)$ and V any closed subspace of $H^1(\Omega)$ containing $H_0^1(\Omega)$, it can be shown (see, e.g., [18, p. 29]) that $a(q)(\cdot, \cdot)$ satisfies (A')-(C'). The operator $A(q)$ is given formally by

$$A(q) = -\sum_{i,j=1}^{l} \frac{\partial}{\partial x_j} a_{ij}(x) \frac{\partial}{\partial x_i} + \sum_{i=1}^{l} b_i(x) \frac{\partial}{\partial x_i} + c(x). \tag{4.4}$$

When $\partial\Omega$ is sufficiently smooth, $A(q)$ is the elliptic operator given by (4.4), V is chosen to be either $H_0^1(\Omega)$ or $H^1(\Omega)$, and (3.1) becomes a parabolic partial differential equation with either Dirichlet or Neumann boundary conditions.

For $H = L_2(\Omega)$ and V a subspace of $H^1(\Omega)$, choosing the approximating subspaces to be the span of an appropriate collection of first-order spline functions will typically satisfy assumption (D) (see [15, Chaps. 2, 6] and Example 4.2 below).

Example 4.2 (Nonlinear elliptic operators). Let Ω be a bounded region in R^l with smooth boundary $\Gamma = \partial\Omega$. For $\alpha = (\alpha_1, \cdots, \alpha_l)$ a multi-index, let $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_l$ and denote the αth order generalized, or distributional derivative of a function u by $D^\alpha u$; that is,

$$D^\alpha u(x) = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_l}}{\partial x_l^{\alpha_l}} u(x), \qquad x \in \Omega.$$

Let m be a nonnegative integer and let δu denote the vector-valued function of length $N = \binom{l+m}{l}$ whose components are all of the partial derivatives of u of order greater than or equal to zero and less than or equal to m.

For each multi-index α with $|\alpha| \leqq m$, let $(x, \zeta) \to a_\alpha(x, \zeta)$ be a real-valued function defined on $\Omega \times R^N$ that is measurable in x and continuous in ζ. We assume that

(1) There exist a $g \in L_2(\Omega)$ and a positive constant γ such that

$$|a_\alpha(x, \zeta)| \leqq \gamma(|p(\zeta)| + g(x)) \quad \text{where } p(\zeta) \text{ is any polynomial,} \tag{4.5}$$

for almost every $x \in \Omega$, each $\zeta \in R^N$, and all α with $|\alpha| \leqq m$; and

(2) There exists a positive constant λ such that

$$\sum_{|\alpha| \leqq m} (a_\alpha(x, \zeta) - a_\alpha(x, \eta))(\zeta_\alpha - \eta_\alpha) \geqq \lambda \sum_{|\alpha| \leqq m} |\zeta_\alpha - \eta_\alpha|^2 \tag{4.6}$$

for almost every $x \in \Omega$ and all $\zeta, \eta \in R^N$.

Let $H = L_2(\Omega)$ and let V be any closed subspace of $H^m(\Omega)$ that contains $H_0^m(\Omega)$. Define the operator $A: V \to V^*$ by

$$(Au)(v) = \sum_{|\alpha| \leqq m} \int_\Omega a_\alpha(x, \delta u(x)) D^\alpha v(x)\, dx, \tag{4.7}$$

for $u, v \in V$. The operator A given by (4.7) is the distributional form of the formal differential operator

$$(Au)(x) = \sum_{|\alpha| \leqq m} (-1)^\alpha D^\alpha a_\alpha(x, \delta u(x)). \tag{4.8}$$

A differential operator of the form (4.8) is referred to as a nonlinear elliptic operator and the partial differential equation

$$\frac{\partial u}{\partial t}(t, x) + \sum_{|\alpha| \leqq m} (-1)^\alpha D^\alpha a_\alpha(x, \delta u(t, x)) = f(t, x) \tag{4.9}$$

is said to be of nonlinear parabolic type. When $V = H_0^m(\Omega)$, a solution in V^* to the abstract equation

$$\dot{u}(t) + Au(t) = f(t),$$

with A given by (4.7), corresponds to a variational solution to (4.9) which satisfies Dirichlet boundary conditions. When $V = H^m(\Omega)$, a variational solution to the Neumann problem is obtained. Note that in the linear case we have

$$a_\alpha(x, \delta u(x)) = \sum_{|\beta| \leqq m} a^{\alpha,\beta}(x) D^\beta u(x).$$

Under the assumptions above, it is not difficult to show that A given by (4.7) is hemicontinuous and satisfies conditions (B) and (C) given in § 3. With an appropriate choice of the space $\mathcal{Q}$ and the set Q, condition (A) can be satisfied as well.

A quasilinear model for heat conduction or mass transfer, in which the heat or mass flux is a function of the temperature or mass fraction gradient discussed in [16] and [17], leads to a nonlinear elliptic operator and a nonlinear parabolic partial differential equation of the forms (4.8) and (4.9), respectively, with $m = 1$. Let Ω be a bounded region in R^l with smooth boundary and let $\mathcal{Q} = L_\infty(\Omega \times R^l)$. Let Q be a compact subset of $\mathcal{Q}$ with the property that $q \in Q$ if and only if the mapping $\zeta \to q(x, \zeta)$ is C^1 on R^l for almost every $x \in \Omega$, and there exists a $\lambda > 0$ (which does not depend on q) such that

$$(q(x, \zeta)\zeta - q(x, \eta)\eta) \cdot (\zeta - \eta) \geqq \lambda |\zeta - \eta|^2 \tag{4.10}$$

for almost every $x \in \Omega$ and all $\zeta, \eta \in R^l$. (When $l = 1$, the function $q(x, \xi) = q(\xi) = (1 - 0.5\, e^{-\xi^2})$ satisfies (4.10).)

Let $H = L_2(\Omega)$ and let V be any closed subspace of $H^1(\Omega)$ that contains $H_0^1(\Omega)$. Then $V \subset H \subset V^*$, and for each $q \in Q$ define $A(q): V \to V^*$ by

$$(A(q)u)(v) = \int_\Omega q(x, \nabla u(x)) \nabla u(x) \cdot \nabla v(x)\, dx \tag{4.11}$$

for $u, v \in V$. Note that for each $q \in Q$ the operator given by (4.11) is of the form (4.7) with

$$a_\alpha(x, \delta u(x)) = q(x, \nabla u(x)) D^\alpha u(x) \tag{4.12}$$

for $x \in \Omega$ and all α with $|\alpha| = 1$ and $a_\alpha = 0$ for $|\alpha| = 0$. The nonlinear parabolic partial differential equation (4.9) takes the form

$$\frac{\partial u}{\partial t}(t, x) - \nabla \cdot q(x, \nabla u(t, x)) \nabla u(t, x) = f(t, x), \qquad t > 0, \quad x \in \Omega.$$

Taking $\| \cdot \|$ to be the usual norm on $H^1(\Omega)$, it follows that

$$\| A(q_0)u - A(q_1)u \|_* \leqq |q_0 - q_1|_{L_\infty} \| u \|$$

for each $u \in V$ and $q_0, q_1 \in Q$. Since Q is a compact subset of $L_\infty(\Omega \times R^l)$, it is easily verified that a_α given by (4.12) satisfies a growth condition of the form (4.5) with p linear and γ and g independent of $q \in Q$. An application of the Mean Value Theorem together with assumption (4.10) implies the existence of a $\lambda > 0$, independent of $q \in Q$, for which (4.6) holds. Consequently, (A), (B), and ($\tilde{\text{C}}$) given in § 3 are satisfied, and our general theory (including Lemma 3.2) can be applied.

With regard to approximation, polynomial spline function based Galerkin subspaces can often be shown to satisfy condition (D). For example, when $l = 1$ and $\Omega = (0, 1)$ in the nonlinear heat conduction/mass transfer example discussed above, the subspaces H_n can be chosen as the span of the linear B-spline ("hat") functions with respect to the uniform mesh $\{0, 1/n, 2/n, \cdots, 1\}$ appropriately modified to satisfy stable, or geometric, boundary conditions. Familiar error estimates for interpolation and the Schmidt inequality can then be used to verify that condition (D) is satisfied (see, e.g., [15, Chap. 6.3]). Generalization to higher dimensions is possible, and can often be achieved via tensor products of one-dimensional elements (again, see [15, Chap. 6]).

5. Concluding remarks. We have developed a general abstract approximation framework for the identification of nonlinear distributed parameter evolution systems. The class of systems to which our theory applies are those whose dynamics can be described by a nonlinear operator that satisfies conditions that are the natural nonlinear extensions, or analogues, of the properties of regularly dissipative, or abstract parabolic, linear operators. The approach we have taken is based on the defining of a sequence of approximating finite-dimensional identification problems in which the systems to be identified are Galerkin approximations to the original, underlying, infinite-dimensional nonlinear dynamics. Under a weak continuity assumption with respect to the unknown parameters to be identified, equiboundedness and equimonotonicity conditions, and an approximation assumption on the Galerkin subspaces (all of which are readily verified for wide classes of nonlinear distributed systems and finite-element subspaces), we are able to demonstrate that solutions to the approximating problems exist, and, in some sense, approximate (i.e., subsequential convergence) solutions to the original infinite-dimensional identification problem. We have shown that the linear theory presented in [2] and [3] is a special case of our nonlinear framework and that our results are applicable to a reasonably wide class of nonlinear elliptic operators and corresponding nonlinear parabolic partial differential equations. In particular, we have considered application of our theoretical framework to a quasilinear model for heat conduction or mass transport.

The general approximation result for nonlinear evolution systems discussed in § 2 is applicable to a much broader class of nonlinear dynamical systems than we subsequently treated in § 3. For example, this class of systems would include those with dynamics described by set-valued maps or multifunctions, and (after nontrivial modification to the general theory) time-dependent or nonautonomous operators. We

are currently investigating these features of the general approximation theory in the context of parameter estimation problems. (In this connection, see the last remark in § 3.) Also, we would like to be able to weaken the somewhat restrictive strong monotonicity condition. Any progress that we might make in these efforts would have the potential to significantly enlarge the class of nonlinear systems to which our theory and framework would apply. Finally, extensive numerical or computational studies designed to demonstrate the feasibility and point out the limitations of our schemes and general approach are currently underway and will be reported on in a forthcoming paper.

After this paper was accepted for publication, related efforts by Kluge and Langmach were called to the authors' attention. Some of these efforts are related in that those authors used monotonicity concepts to treat several specific estimation problems for nonlinear partial differential equations. Motivated by problems involving flow of viscous liquids through porous media [H. Langmach, *On the determination of functional parameters in some parabolic differential equations*, in Theory of Nonlinear Operators (Proc. Summer School Berlin 1977), Abh. Akad. Wiss. DDR, 6 (1978), pp. 174-184], Langmach discusses conditions to guarantee existence and convergence of approximations to best fit parameters in first-order parabolic systems where the unknown coefficients are monotone functions of the gradient of the system solution. His approach is in the spirit of the efforts on inverse problems for elliptic systems by Kluge and Langmach [*On some problems of determination of functional parameters in partial differential equations*, in Modeling and Identification of Distributed Parameter Systems, Lecture Notes in Control and Information Science 1, Springer-Verlag, Berlin, New York, 1977, pp. 298-309] and differs substantially from the approach taken in §§ 2 and 3 of this paper.

REFERENCES

[1] H. T. BANKS, S. S. GATES, I. G. ROSEN, AND Y. WANG, *The identification of a distributed parameter model for a flexible structure*, SIAM J. Control Optim., 26 (1988), pp. 743-762.

[2] H. T. BANKS AND K. ITO, *A theoretical framework for convergence and continuous dependence of estimates in inverse problems for distributed parameter systems*, Appl. Math. Lett., 1 (1988), pp. 13-17.

[3] ———, *A unified framework for approximation and inverse problems for distributed parameter systems*, Control-Theory Adv. Tech., 4 (1988), pp. 73-90.

[4] H. T. BANKS AND I. G. ROSEN, *Computational methods for the identification of spatially varying stiffness and damping in beams*, Control-Theory Adv. Tech., 3 (1987), pp. 1-32.

[5] ———, *Numerical schemes for the estimation of functional parameters in distributed models for mixing mechanisms in lake and sea sediment cores*, Inverse Problems, 3 (1987), pp. 1-23.

[6] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff International, Leyden, the Netherlands, 1976.

[7] M. G. CRANDALL AND L. C. EVANS, *On the relation of the operator $\partial/\partial s + \partial/\partial t$ to evolution governed by accretive operators*, Israel J. Math., 21 (1975), pp. 261-278.

[8] M. G. CRANDALL AND A. PAZY, *Nonlinear evolution equations in Banach space*, Israel J. Math., 11 (1972), pp. 57-94.

[9] L. C. EVANS, *Nonlinear evolution equations in an arbitrary Banach space*, Israel J. Math., 26 (1977), pp. 1-42.

[10] J. A. GOLDSTEIN, *Approximation of nonlinear semigroups and evolution equations*, J. Math. Soc. Japan, 24 (1972), pp. 558-573.

[11] ———, *Semigroups of Linear Operators and Applications*, Oxford, New York, 1985.

[12] J. K. HALE, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[13] J. KISYNSKI, *A proof of the Trotter-Kato theorem on approximation of semigroups*, Colloq. Math., 18 (1967), pp. 181-184.

[14] S. REICH, *Convergence and approximation of nonlinear semigroups*, J. Math. Anal. Appl., 76 (1980), pp. 77-83.

[15] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
[16] J. C. SLATTERY, *Quasi-linear heat and mass transfer*, I. *The constitutive equations*, Appl. Sci. Res., 12, Sec. A (1963), pp. 51-56.
[17] ———, *Quasi-linear heat and mass transfer*, II. *Analyses of experiments*, Appl. Sci. Res., 12, Sec. A (1963), pp. 57-65.
[18] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 570-581, May 1990

006

OBSERVABILITY OF SYSTEMS ON LIE GROUPS AND COSET SPACES*

D. CHENG†, W. P. DAYAWANSA‡, AND C. F. MARTIN§

Abstract. The purpose of this paper is to study the observability of a class of systems for which the state space is a Lie group and the output space is a coset space. The study of this type of system was initiated by Brockett [*SIAM J. Control*, 10 (1972), pp. 265-284]. In this paper, Brockett's observability results are generalized and necessary and sufficient conditions for observability are obtained. Effective algorithms are established to verify such conditions. Finally, as an application, some disturbance decoupling problems are considered.

Key words. observability, global observability, Lie groups, analytic systems

AMS(MOS) subject classification. 93B

1. Introduction. In this paper, we study the observability properties of systems that are described by a state equation that evolves on a Lie group G and an output equation that takes values in a coset space of G. These equations are assumed to be of the form

$$\dot{x} = A(x) + \sum_{i=1}^{m} B_i(x)u_i, \qquad x \in G, \tag{1.1}$$

$$y = Cx \tag{1.2}$$

where $A(x)$, $B_1(x), \cdots, B_m(x)$ are right-invariant vector fields on G, C is a closed subgroup of G, and the notation Cx is to be interpreted as the right coset of C in G that contains x.

This system model has been studied by Brockett [1] where G was assumed to be a group of matrices. Brockett has shown [1] that there are many important applications in engineering and in physics that have models of this form. Jurdjevic and Sussmann [2] have extended (1.1) to an abstract Lie group G and have obtained a set of basic controllability properties of (1.1). Our work is related to and extends the work of [1].

The observability properties are discussed by Brockett [1]. To describe Brockett's observability result, we need a preliminary definition.

DEFINITION 1.1. Two points x_1 and x_2 are distinguishable if there exists some control that gives rise to different outputs for the two starting points.

Let S be a subset of G. We denote by $\{S\}_G$ the subgroup generated by S, i.e., the smallest subgroup of G containing S. Let $\mathcal{H}$ be a set of right invariant vector fields of G. We denote by $\{\mathcal{H}\}_{\mathrm{LA}}$ the Lie subalgebra generated by $\mathcal{H}$, and

$$\exp(\{\mathcal{H}\}_{\mathrm{LA}}) = \{\exp X \mid X \in \{\mathcal{H}\}_{\mathrm{LA}}\}.$$

The main observability result of [1] is Theorem 1.1.

* Received by the editors May 31, 1988; accepted for publication (in revised form) July 12, 1989.

† Department of Mathematics, Texas Tech University, Lubbock, Texas 79409.

‡ Department of Mathematics, Texas Tech University, Lubbock, Texas 79409. This work was supported in part by National Science Foundation grant ECS-88024831.

§ Department of Mathematics, Texas Tech University, Lubbock, Texas 79409. This work was supported in part by National Security Agency grant MDA 904-85-H0009 and National Aeronautics and Space Administration grant NAG 2-82.

THEOREM 1.1 [1]. *Let $\mathscr{H}$ and $\mathscr{L}$ be Lie algebras in* $\mathrm{gl}\{n, \mathbb{R}\}$, *and suppose that all the points reachable from the identity for matrix system*

$$\dot{x} = \left(A + \sum_{i=1}^{m} u_i B_i\right)x, \qquad y = (\{\exp \mathscr{H}\}_G)x$$

are $\{\exp \mathscr{L}\}_G$. *Then the set of initial states $\mathscr{S}$, which are indistinguishable from the identity, contains* $\{\exp \mathscr{K}\}_G$ *if and only if* $\{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_{\mathrm{LA}} \subset \mathscr{H}$. *Therefore a necessary condition for all states to be distinguishable from the identity is that $\mathscr{H}$ contains no subalgebra $\mathscr{K}$ such that* $\{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_{\mathrm{LA}} \subset \mathscr{H}$.

It is shown by example in Brockett [1] that the preceding theorem is not sufficient. An important point in this theorem is that the "unobservable" part is related to an $\mathscr{L}$-invariant subalgebra $\{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_{\mathrm{LA}}$, which is contained in the Lie subalgebra of the output subgroup.

Motivated by this fact, we investigate the "unobservable" part in more detail. The results of Jurdjevic and Sussmann [2] enable us to describe the controllable set, which corresponds to $\{\exp \mathscr{L}\}_G$ of Theorem 1.1. Based on [1] and [2], we give necessary and sufficient conditions for the system (1.1), (1.2) to be observable.

The paper is organized in the following way. Section 2 contains two main results—local observability conditions and global observability conditions. In § 3, we develop algorithms that are useful for studying groups of matrices. In § 4, we give some examples. Finally, in § 5 the input-output decoupling problem is discussed as an application.

2. Observability results. To avoid unnecessary complexity, we assume throughout this paper that the controls are piecewise constant. In fact, this is not essential. For instance, if we replace the set of piecewise constant functions by the set of the piecewise continuous functions, all of the arguments remain valid.

Let $R(x)$ be the reachable set starting from x, i.e., $R(x)$ is the set of points y such that there exist a piecewise constant control u and a time $T \geqq 0$, such that the solution of (1.1) satisfies $x(0) = x$, $x(T) = y$. We denote by $R(x, t)$ the reachable set at time t, starting from x.

It is proved in Jurdjevic and Sussmann [2] that for the right-invariant system (1.1), the reachable set of x is related to the reachable set of the identity e by

$$R(x) = R(e)x. \tag{2.1}$$

Using this fact, we prove the following elementary result, which shows that distinguishing two arbitrary points is equivalent to distinguishing a point from the identity.

LEMMA 2.1. *Two points p and q are indistinguishable if and only if for each* $r \in R(e)$

$$\mathrm{Ad}\,(r)pq^{-1} \in C. \tag{2.2}$$

Proof. By the structure of the output (1.2) it is clear that p and q are indistinguishable if and only if for all t, $R(p, t)$ and $R(q, t)$ are in the same coset of C. From (2.1), it follows that

$$Crp = Crq \quad \text{for all } r \in R(e), \tag{2.3}$$

that is,

$$rpq^{-1}r^{-1} = \mathrm{Ad}\,(r)pq^{-1} \in C. \qquad \square$$

Now we may define an unobservable state as follows. (It is similar to the linear case: x_1 and x_2 are indistinguishable if and only if $x_1 - x_2$ belongs to an unobservable subspace.)

DEFINITION 2.1. A point h is called unobservable if there exist p and q such that $pq^{-1} = h$ and p and q are indistinguishable.

Remark 1. Let h be unobservable. Then it follows from Lemma 2.1 that for any pair (p^1, q^1) if $p^1(q^1)^{-1} = h$, then p^1 and q^1 are indistinguishable.

Let

$$H = \{h \in G \mid h \text{ is unobservable}\}.$$

By definition of unobservable state and equation (2.2), it is clear that

$$H \subset C. \tag{2.4}$$

In fact, H has a subgroup structure that is shown in the following lemma.

LEMMA 2.2. *Assume C is closed. Then the unobservable set H is a closed Lie subgroup of G.*

Proof. By definition and Lemma 2.1,

$$H = \{h \in G \mid rhr^{-1} \in C \text{ for all } r \in R(e)\}. \tag{2.5}$$

Let $h_1, h_2 \in H$. Then,

$$rh_1h_2^{-1}r^{-1} = rh_1r^{-1}rh_2^{-1}r^{-1} = (rh_1r^{-1})(rh_2r^{-1})^{-1} \in C.$$

Thus, H is a subgroup of G.

Since C is closed, if for a sequence $\{h_n\} \subset H$, $h_n \to h$, as $n \to \infty$, then

$$rh_nr^{-1} \to rhr^{-1} \in C.$$

Thus, $h \in H$, and hence H is closed. Now the result follows from the well-known fact (see for example, Hausner and Schwartz [4]) that a closed subgroup of a Lie group is a Lie subgroup. □

If C is closed, the output mapping has an analytic structure that is described by the following well-known theorem.

THEOREM 2.1 [3]. *Let G be a Lie group and C a closed subgroup of G. Then the quotient space $C\backslash G$ admits the structure of real analytic manifold in such a way that the action of G on $C\backslash G$ is real analytic, that is, the mapping $G \times C\backslash G \to C\backslash G$, which maps (p, Cq) into Cpq, is real analytic. In particular, the projection $G \to C\backslash G$ is real analytic.*

Let $\{R(e)\}_G$ be the subgroup of G generated by $R(e)$ and let $\overline{\{R(e)\}_G}$ denote the closure of $\{R(e)\}_G$. For convenience denote the vector fields $A(x), B_1(x), \cdots, B_m(x)$ by $A, B_1, \cdots, B_m$, respectively, where A and B_i are elements in $\mathcal{G}(G)$, the Lie algebra of G. Then we have the following lemma.

LEMMA 2.3. *Assume $h \in H$. Then*

$$\text{Ad}\,(r)h \in H \quad \textit{for all } r \in \overline{\{G(e)\}_G}. \tag{2.6}$$

Proof. First, we claim that

$$\text{Ad}\,(r)h \in H \quad \text{for all } r \in G(e). \tag{2.7}$$

Since $R(e)$ is a semigroup [2], for any $\tilde{r} \in R(e)$ we have $\tilde{r}r \in R(e)$. Thus,

$$(\tilde{r}r)h(\tilde{r}r)^{-1} = \tilde{r}(rhr^{-1})\tilde{r}^{-1} \in C \quad \text{for all } \tilde{r} \in R(e).$$

It follows that $rhr^{-1} \in H$.

From its defining properties, it is clear that

$$\{R(e)\}_G = \left\{ \exp(t_sX_s) \cdots \exp(t_1X_1) \mid t_i \in \mathbb{R}, s \in Z^+, \right.$$
$$\left. X_i \in \left\{ A + \sum_{j=1}^{m} u_jB_j \;\middle|\; u_j \in \mathbb{R} \right\}, i = 1, \cdots, s \right\}. \tag{2.8}$$

Set

$$E_s = \{(t_1, \cdots, t_s) \in \mathbb{R}^s \,|\, \mathrm{Ad}\,(\exp(t_s X_s) \cdots \exp(t_1 X_1))h \in C\}.$$

Then, to prove (2.6) for $r \in \{R(e)\}_G$ it is enough to show that $E_s = \mathbb{R}^s$, $s = 1, 2, \cdots$. We proceed by induction. For $s = 1$, if $\mathrm{Ad}\,(\exp t_1 X_1)h \not\subset C$, then there exists $\tilde{t}_1$ such that

$$\left.\frac{d}{dt_1}\right|_{\tilde{t}_1} \mathrm{Ad}\,(\exp t_1 X_1)h \notin (R_p)_* \mathcal{G}(C)$$

where $p = \mathrm{Ad}\,(\exp \tilde{t}_1 X_1)h$, $\mathcal{G}(C)$ is the Lie algebra of C and R_p is the right translation, i.e., $R_p : G \to G$ is defined as $x \to xp$. In other words, there exists a right-invariant one-form $w(x)$ generated by $w \in (\mathcal{G}(C))^{\perp}$ such that

$$\left\langle w_{(p)}, \left.\frac{d}{dt_1}\right|_{\tilde{t}_1} \mathrm{Ad}\,(\exp t_1 X_1)h \right\rangle \neq 0. \tag{2.9}$$

By analyticity, (2.9) holds in an open dense subset of $\mathbb{R}$. But according to (2.7), for $\tilde{t}_1 \in \mathbb{R}_+ = \{t \in \mathbb{R},\ t \geqq 0\}$ the left-hand side of (2.9) is zero; this leads to a contradiction. Now, assume that

$$\mathrm{Ad}\,(\exp(t_{s-1} X_{s-1}) \cdots \exp(t_1 X_1))h \in C, \qquad t_i \in \mathbb{R}$$

and

$$\{\mathrm{Ad}\,(\exp(t_s X_s) \cdots \exp(t_1 X_1))h \,|\, (t_1, \cdots, t_s) \in \mathbb{R}^s\} \not\subset C.$$

Then there exists $\tilde{t} = (\tilde{t}_1, \cdots, \tilde{t}_s)$ such that

$$\left.\frac{d}{dt}\right|_{\tilde{t}_s} \mathrm{Ad}\,(\exp(tX_s) \exp(\tilde{t}_{s-1} X_{s-1}) \cdots \exp(\tilde{t}_1 X_1))h \in \mathcal{G}(C).$$

Similar to the case when $s = 1$, we have a contradiction.

Thus, we have shown that (2.6) holds for all $r \in \{R(e)\}_G$. By continuity, it holds for all $r \in \overline{\{R(e)\}_G}$. □

Remark 2.[1] It is clear by (2.8) that $\{R(e)\}_G$ is a path-connected group, hence a Lie subgroup [5]. Now since $\{R(e)\}_G$ is a connected Lie group, and $A, B_1, B_2, \cdots, B_m$ generate $\mathcal{G}(\{R(e)\}_G)$, then [2, Lemma 6.2]

$$\begin{aligned} \{R(e)\}_G = \{&\exp(t_x X_s) \cdots \exp(t_1 X_1) \,|\, t_i \in \mathbb{R},\ s \in Z^+, \\ &X_i \in \{A, B_1, \cdots, B_m\},\ i = 1, \cdots, s\}. \end{aligned} \tag{2.10}$$

□

Next, we investigate the relations among the Lie algebras $\mathcal{G}(H)$, $\mathcal{G}(C)$, and $\mathcal{G}(G)$, which are the Lie algebras of H, C, and G, respectively.

Let $\{X_1(x), \cdots, X_s(x)\}$ be a set of right-invariant vector fields generated by $X_i \in \mathcal{G}(G)$, $i = 1, \cdots, s$, respectively. Let Δ denote the subspace of $\mathcal{G}(G)$ spanned by $\{x_1, \cdots, x_5\}$. A subspace Δ of $\mathcal{G}(G)$ is called $Y \in \mathcal{G}(G)$ invariant if

$$\{[Y, X] \,|\, X \in \Delta\} \subset \Delta.$$

Likewise, for right-invariant we form $w_1(x), \cdots, w_s(x)$ generated by $w_i \in \mathcal{G}^*(G)$, the cotangent space of G at the identity e, we have a right-invariant subspace

$$\Omega = \mathrm{span}\,\{w_1, \cdots, w_s\}.$$

[1] This remark was suggested by an anonymous reviewer.

Ω is Y invariant if

$$\{L_Y w \mid w \in \Omega\} \subset \Omega.$$

The following two lemmas are generalizations of Theorem 1.1.

LEMMA 2.4. *$\mathcal{G}(H)$ is A and B_i, $i = 1, \cdots, m$, invariant.*

Proof. Let $X = A$ or B_i, $t \in \mathbb{R}$, $p = \exp(tX)$. According to Lemma 2.3 and (2.10), $(\operatorname{Ad} \exp(tX))_* \mathcal{G}(H) \subset \mathcal{G}(H)$. Now let $Y \in \mathcal{G}(H)$. Then

$$[X, Y] = \frac{d}{dt}\bigg|_{t=0} \operatorname{Ad} \exp(tX)_* Y \in \mathcal{G}(H). \qquad \square$$

LEMMA 2.5. *$\mathcal{G}(H)$ is the largest A and B_i, $i = 1, \cdots, m$, invariant Lie subalgebra contained in $\mathcal{G}(C)$.*

Proof. We claim that

$$\mathcal{G}(H) = \bigcap_{\substack{X_1, \cdots, X_p \in \{A, B_1, \cdots, B_m\} \\ p \in Z^+}} \operatorname{ad}_{X_1}^{-1} \cdots \operatorname{ad}_{X_p}^{-1} \mathcal{G}(C). \tag{2.11}$$

First, we show that (2.11) implies $\mathcal{G}(H)$ is the largest A and B_i invariant Lie subalgebra contained in $\mathcal{G}(C)$. Assume $\mathcal{G}(\tilde{H}) \subset \mathcal{G}(C)$ is also A and B_i invariant. Then, for any $X_1, \cdots, X_p \in \{A, B_1, \cdots, B_m\}$,

$$\operatorname{ad}_{X_1} \cdots \operatorname{ad}_{X_p} \mathcal{G}(\tilde{H}) \subset \mathcal{G}(\tilde{H}) \subset \mathcal{G}(C).$$

Thus,

$$\mathcal{G}(\tilde{H}) \subset \operatorname{ad}_{X_1}^{-1} \cdots \operatorname{ad}_{X_p}^{-1} \mathcal{G}(C).$$

Since $X_1, \cdots, X_p$ are chosen arbitrarily, we have that

$$\mathcal{G}(\tilde{H}) \subset \mathcal{G}(H).$$

Next, we prove (2.11).

($\subseteq$) Lemma 2.4 shows that $\mathcal{G}(H)$ is A and B_i invariant. The inclusion follows by an argument similar to the above.

($\supseteq$) Let

$$Y \in \bigcap_{\substack{X_1, \cdots, X_p \in \{A, B_1, \cdots, B_m\} \\ P \in Z^+}} \operatorname{ad}_{X_1}^{-1} \cdots \operatorname{ad}_{X_p}^{-1} \mathcal{G}(C).$$

To show that $Y \in \mathcal{G}(H)$, it is enough to show that

$$\exp(\tau Y) \in H \quad \text{for all } \tau \in \mathbb{R}.$$

Using (2.10), it suffices to show that for any

$$\begin{aligned} &X_1, \cdots, X_p \in \{A, B_1, \cdots, B_m\}, \qquad (t_1, \cdots, t_p) \in \mathbb{R}^p, \\ &\operatorname{Ad}(\exp(t_p X_p) \cdots \exp(t_1 X_1)) \exp \tau Y \in C. \end{aligned} \tag{2.12}$$

Since $\operatorname{Ad}(\exp(t_p X_p) \cdots \exp(t_1 X_1))$ is a diffeomorphism, we have

$$\operatorname{Ad}(\exp(t_p X_p) \cdots \exp(t_1 X_1)) \exp \tau Y = \exp(\operatorname{Ad}(\exp(t_p X_p) \cdots \exp(t_1 X_1) \tau Y)).$$

Now to prove (2.12), it suffices to show that

$$\operatorname{Ad}(\exp(t_p X_p) \cdots \exp(t_1 X_1)) \tau Y \in \mathcal{G}(C). \tag{2.13}$$

Let us denote the right-hand side of (2.11) by $\mathcal{J}$. Now since

$$\begin{aligned} &\operatorname{Ad}(\exp(t_p X_p) \cdots \exp(t_1 X_1) \tau Y) \\ &\qquad = \operatorname{Ad}(\exp t_p X_p) \operatorname{Ad}(\exp t_{p-1} X_{p-1}) \cdots \operatorname{Ad}(\exp(t_1 X_1)) \tau Y, \end{aligned}$$

it suffices to prove that

$$\operatorname{Ad}(\exp(t_1X_1)\tau Y)\in\mathscr{J}.$$

But

$$\operatorname{Ad}(\exp(t_1x_1)\tau y)=\exp(\operatorname{ad}(t_1x_1))\tau y$$
$$=\sum_{n=0}^{\infty}\frac{(\operatorname{ad}(t_1x_1))^n}{n!}(\tau y).$$

Therefore, obviously, $\operatorname{Ad}(\exp(t_1x_1)\tau y)\in\mathscr{J}$. □

We are now ready to discuss the observability properties of (1.1), (1.2).

DEFINITION 2.2. System (1.1), (1.2) is locally observable at x if there exists a neighborhood V_x of x such that

$$I_x\cap V_x=\{x\},$$

where I_x is the set of points that are indistinguishable from x. System (1.1), (1.2) is locally observable if it is locally observable everywhere. System (1.1), (1.2) is (globally) observable if

$$I_x=\{x\}.$$

In fact, Lemma 2.5 leads to the following local observability result, which is now obvious.

THEOREM 2.2. *System* (1.1), (1.2) *is locally observable if and only if the largest* A *and* B_i, $i=1,\cdots,m$, *invariant subalgebra contained in* $\mathscr{G}(C)$ *is zero. Moreover, if* V_e *is a neighborhood of* e *such that* $I_e\cap V_e=\{e\}$, *then* $V_x=R_x(V_e)$ *is a neighborhood of* x *such that* $I_x\cap V_x=\{x\}$.

Let S be the centralizer of $\{R(e)\}_G$, i.e.,

$$S=\{x\in G\,|\,rx=xr \text{ for all } r\in\{R(e)\}_G\}. \tag{2.14}$$

According to (2.10), we may express S in an easily verifiable form as

$$S=\{x\in G\,|\,x\exp(tX)=\exp(tX)x,\ X\in\{A,B_1,\cdots,B_m\}\}. \tag{2.15}$$

We will use S to establish a global result.

It is obvious that S is a closed subgroup of G, and hence is a Lie subgroup. Moreover, by the construction of $\{R(e)\}_G$ we see that to verify that $x\in S$ it is enough to verify that

$$\operatorname{Ad}(x)\exp(tY)=\exp tY$$

for

$$Y\in\{A,B_1,\cdots,B_m\},\qquad t\in\mathbb{R}.$$

Now we state our global observability theorem.

THEOREM 2.3. *System* (1.1), (1.2) *is globally observable if and only if the following two conditions are satisfied*:

(a) $\mathscr{G}(H)=\{0\}$,

(b) $S\cap C=\{e\}$.

Proof. Necessity. The necessity of (a) has been proved in Theorem 2.2. The necessity of (b) is obvious, because if $e\neq h\in S\cap C$, then $h\in H$, i.e., h is indistinguishable from e.

Sufficiency. From (a) we see that H is a discrete subgroup of G. Now for each $h \in H$, we define a mapping $\phi : \{R(e)\}_G \to H$ as

$$\phi(r) = \mathrm{Ad}\,(r)h.$$

According to Lemma 2.3, ϕ maps $\{R(e)\}_G$ into H. Now, since $\{R(e)\}_G$ is connected and ϕ is continuous, $\{\mathrm{Ad}\,(r)h \,|\, r \in \{R(e)\}_G\} \subset H$ is connected, but since

$$h \in \{\mathrm{Ad}\,(r)h \,|\, r \in \{R(e)\}_G\}$$

it follows that

$$\{h\} = \{\mathrm{Ad}\,(r)h \,|\, r \in \{R(e)\}_G\},$$

i.e.,

$$\mathrm{Ad}\,(r)h = h \quad \text{for all } r \in \{R(e)\}_G.$$

In other words, $h \in S$. Now using condition (b), we see that $h = e$, i.e., $H = \{e\}$. □

3. Algorithm. In the previous section, we saw that the Lie subalgebra $\mathcal{G}(H)$ of the unobservable Lie group H plays an important role in investigating the observability of the system (1.1), (1.2). The following algorithm gives a method to compute it.

ALGORITHM 3.1.

$$\Omega_0 \triangleq \mathcal{G}(C)^{\perp},$$

$$\Omega_{k+1} \triangleq \Omega_k + L_A\Omega_k + \sum_{i=1}^{m} L_{B_i}\Omega_k, \qquad k \geqq 1.$$

Algorithm 3.1 produces an increasing sequence of right-invariant subspaces. To see that it provides $\mathcal{G}(H)$, we need the following theorem. The proof may be found in Isidori [6].

THEOREM 3.1. *In Algorithm* 3.1 *if* $\Omega_{k^*+1} = \Omega_{k^*}$ *then*

$$\mathcal{G}(H) = \Omega_{k^*}^{\perp}. \tag{3.1}$$

Note that the algorithm converges since the sequence of subspace $\{\Omega_K\}$ is increasing.

Since every Lie algebra over the field of real numbers $\mathbb{R}$ is isomorphic to some matrix algebra, we may consider further algorithmic details for the Lie algebras of groups of matrices.

First, let $\omega(x) \in V^*(G)$ be a right-invariant covector field (one-form) generated by $\omega \in \mathcal{G}^*(G)$, and let $A(x)$, $B(x) \in V(G)$ be the right-invariant vector fields generated by A, $B \in \mathcal{G}(G)$, respectively. Then,

$$\begin{aligned} \langle L_A\omega, B\rangle &= \langle L_{A(x)}\omega(x), B(x)\rangle \\ &= L_{A(x)}\langle \omega(x), B(x)\rangle - \langle \omega(x), [A(x), B(x)]\rangle. \end{aligned} \tag{3.2}$$

Since $\langle \omega(x), B(x)\rangle$ is constant, the first term of the right-hand side of (3.2) is zero. Thus, we have

$$\langle L_A\omega, B\rangle = -\langle \omega, [A, B]\rangle. \tag{3.3}$$

Now we consider a group of matrices. Assume the group considered is $\mathrm{GL}\,(n, \mathbb{R})$ (or a subgroup of it). Then, A, $B \in \mathrm{gl}\,(n, \mathbb{R})$ may be considered as matrices $A = (a_{ij})$

and $B = (b_{ij})$, respectively. Let $\omega \in \mathrm{gl}^*(n, \mathbb{R})$. We may assume ω is also expressed as a matrix $\omega = (\omega_{ij})$ and define

$$\langle \omega, A \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{ij} a_{ij}. \tag{3.4}$$

Now,

$$\begin{aligned} \langle L_A \omega, B \rangle &= -\langle \omega, [A, B] \rangle \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} (a_{kj} \omega_{ij} b_{ik} - a_{ik} \omega_{ij} b_{kj}) \\ &= \sum_{p=1}^{n} \sum_{q=1}^{n} \left(\sum_{k=1}^{n} \omega_{pk} a_{qk} - a_{kp} \omega_{kq} \right) b_{pq}. \end{aligned}$$

Thus

$$L_A \omega = [\omega, A^T] = \omega A^T - A^T \omega, \tag{3.5}$$

where T stands for transpose. To apply Algorithm 3.1, formula (3.5) is helpful.

Remark 3. As shown in Brockett [1], a right-invariant vector field on $\mathcal{G}(n, \mathbb{R})$ may be written as

$$A(x) = Ax,$$

where $A = A(e)$ and $A(x) = (R_x)_* A(e) = Ax$. Similarly, a right-invariant covector field may be written as

$$\omega(x) = \omega(x^T)^{-1}$$

where $\omega = \omega(e)$ and $\omega(x) = (R_{x^{-1}})^* \omega(e) = \omega(x^T)^{-1}$.

To see this, we only have to show that

$$\langle \omega(x), A(x) \rangle = \langle \omega, A \rangle.$$

In fact, if we denote $y = x^{-1}$, $x = (x_{ij})$, and $y = (y_{ij})$, then

$$\begin{aligned} \langle \omega(x), A(x) \rangle &= \sum_i \sum_j \left(\sum_p \omega_{ip} y_{jp} \right) \left(\sum_q a_{iq} x_{qj} \right) \\ &= \sum_i \sum_p \sum_q \omega_{ip} a_{iq} \left(\sum_j x_{qj} y_{jp} \right) \\ &= \sum_i \sum_p \sum_q \omega_{ip} a_{iq} \delta_{qp} \\ &= \sum_i \sum_p \omega_{ip} a_{ip} = \langle \omega, A \rangle. \end{aligned}$$

In fact, if we rewrite $A(x)$ in the "usual fashion" as a vector

$$A(e) = (a_{11}, \cdots, a_{1n}, a_{21}, \cdots, a_{2n}, \cdots, a_{n1}, \cdots, a_{nn})^T,$$

then

$$A(x) = \underbrace{(x^T \dotplus x^T \dotplus \cdots \dotplus x^T)}_{n \text{ terms.}} A(e)$$

Similarly,

$$\omega(x) = \omega(e)((x^T)^{-1} \dotplus (x^T)^{-1} \dotplus \cdots \dotplus (x^T)^{-1})$$

where "$\dotplus$" denotes the direct sum of matrices, and $(x^T \dotplus x^T \dotplus \cdots \dotplus x^T)$ and $((x^T)^{-1} \dotplus (x^T)^{-1} \dotplus \cdots \dotplus (x^T)^{-1})$ are the Jacobian matrices of R_x and $R_{x^{-1}}$, respectively.

4. Examples. In this section, we present some examples to demonstrate our results and algorithms.

Example 4.1. Consider a system

$$\dot{x} = uBx, \tag{4.1}$$

$$y = Cx \tag{4.2}$$

where $x \in \mathrm{GL}(3, \mathbb{R})$, $C = \mathrm{SO}(3)$, and

$$B = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Now $\mathcal{G}(C)$ is the following set of skew-symmetric matrices:

$$\mathcal{G}(C) = \mathrm{span}\left\{\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}\right\}.$$

According to Algorithm 3.1, we set

$$\Omega_0 = \mathrm{span}\left\{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}\right\}$$

$$\triangleq \mathrm{span}\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}.$$

Using formula (3.5), we see that

$$L_B\omega_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad L_B\omega_5 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad L_B\omega_6 = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Thus, $\Omega_1 = \mathcal{G}^*(G)$ and $k_* = 1$. Therefore,

$$\mathcal{G}(H) = \Omega_1^{\perp} = \{0\}.$$

Next, let us consider

$$S \cap C = \{x \in C \mid x \exp tB = \exp(tB)x, \text{ for all } t \in \mathbb{R}\}.$$

Let $x = (x_{ij}) \in C$. Since

$$\exp tB = \begin{bmatrix} 1 & 0 & 0 \\ t & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

we set

$$x \exp tB = \begin{bmatrix} x_{11} + x_{12} & x_{12} & x_{13} \\ x_{21} + tx_{22} & x_{22} & x_{23} \\ x_{31} + tx_{32} & x_{32} & x_{33} \end{bmatrix} = \exp(tB)x = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ tx_{11} + x_{21} & tx_{12} + x_{22} & tx_{13} + x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}.$$

It follows that

$$x_{12} = 0, \quad x_{11} = x_{22}, \quad x_{13} = 0, \quad x_{32} = 0. \tag{4.3}$$

Since $x \in C = \mathrm{SO}(3)$, the only solutions of (4.3) are

$$x_1 = e, \qquad x_2 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{4.4}$$

According to Theorem 2.3, system (4.1), (4.2) is not globally observable. □

Example 4.2. Consider the following system:

$$\dot{x} = Ax + uBx, \tag{4.5}$$

$$y = Cx \tag{4.6}$$

where B and C are as in Example 4.1, and

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

As in the previous example, we see that $\mathscr{G}(H) = \{0\}$. So the system is locally observable. Now

$$e^{At} = \begin{bmatrix} 1 & 0 & t \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

According to (2.15), we have only to check the commutativity of both x_1 and x_2 of (4.4) with $\exp(tA)$. For x_2 the answer is "no." Therefore, $x_1 = e = I_3$ is the only element in $S \cap C$. It follows that system (4.5), (4.6) is globally observable.

Remark 4. In Example 4.2, if we consider e^{At}, e^{Bt}, $e^{-Bt}e^{-At}e^{Bt}e^{At}$ and their products, it is easy to see that

$$\{R_{(e)}\}_G = \left\{ \begin{pmatrix} 1 & 0 & a \\ b & 1 & c \\ 0 & 0 & 1 \end{pmatrix} \middle| \, a, b, c \in \mathbb{R} \right\}.$$

It follows that

$$S = \left\{ \begin{pmatrix} x & 0 & 0 \\ 0 & x & y \\ 0 & 0 & x \end{pmatrix} \middle| \, x, y \in \mathbb{R}, x \neq 0 \right\}$$

and therefore,

$$S \cap C = I_3.$$

But in general, it is difficult to calculate $\{R(e)\}_G$ and S. In fact, Example 4.2 shows that to use Theorem 2.3 it is not necessary to construct $\{R(e)\}_G$ and S directly. We may check the global observability by the following rule, which may be considered as a corollary of Theorem 2.3.

COROLLARY 4.1. *System* (1.1), (1.2) *is globally observable, if and only if,*

(a) $\mathscr{G}(H) = \{0\}$,

(b) $\{x \in C \mid \exp(tX)x = x \exp(tX),\ X \in \{A, B_1, \cdots, B_m\},\ t \in \mathbb{R}\} = \{e\}$.

5. Decoupling problems. As an application, we consider a decoupling problem. To keep the right-invariance of $A(x)$ and $B_i(x)$, we consider only a constant feedback

$$u = \alpha + \beta u \tag{5.1}$$

where $\alpha \in \mathbb{R}^m$ and $\beta \in \mathrm{GL}(m, \mathbb{R})$.

Now assume

$$\dot{x} = A(x) + \sum_{i=1}^{m} u_i B_i(x) + \omega W(x), \tag{5.2}$$

$$y = Cx \tag{5.3}$$

where ω is a disturbance.

LEMMA 5.1. *The disturbance ω does not affect the output y if and only if*

$$W \in \mathcal{G}(H). \tag{5.4}$$

Proof. In fact, we may choose a local coordinate chart (ϕ, U) around e, say $x = (x^1, x^2)$, such that

$$C \cap U = \{p \in U \mid x_p^2 = 0\}.$$

Thus,

$$y = x^2(q), \qquad q \in U.$$

Now, it is easy to see that on U, $\mathcal{G}(H)$ is the largest A and B_i invariant distribution contained in the $\ker(y_*)$. Note that constant feedback does not affect $\mathcal{G}(H)$. Thus, the canonical decoupling result shows that (Isidori et al. [7]) (5.4) is a necessary and sufficient condition that ω does not affect y on V. By the analyticity, it is also true globally. □

Next, we consider the input-output decoupling problem. Assume $C_1, \cdots, C_k$ are Lie subgroups of G. Let $C = C_1 \cap \cdots \cap C_k$. Then it is easy to see that

$$y = Cx \tag{5.5}$$

is equivalent to

$$\begin{aligned} y_1 &= C_1 x \\ &\vdots \\ y_k &= C_k x \end{aligned} \tag{5.6}$$

in the sense that p and q are indistinguishable in (5.5) if and only if they are indistinguishable in (5.6).

Let $\mathcal{G}(H^i)$ be the largest A and B_i invariant Lie subalgebra contained in $\mathcal{G}(C_1 \cap \cdots C_{i-1} \cap C_{i+1} \cdots \cap C_k)$. Consider the system

$$\begin{aligned} \dot{x} &= A(x) + \sum_{i=1}^{m} u_i B_i(x), \\ y_j &= C_j x, \qquad j = 1, \cdots, k. \end{aligned} \tag{5.7}$$

We say that the input-output decoupling problem is solvable if there exists $\beta = (\beta_{ij}) \in \mathrm{GL}(m, \mathbb{R})$, such that for

$$u = v\beta$$

there exists a partition of v, namely $v = (v^1, \cdots, v^k)$, such that v^i affects only the corresponding y_i, $i = 1, \cdots, k$.

THEOREM 5.1. *For the system described by* (5.7) *the input-output decoupling problem is solvable if and only if*

$$B = B \cap \mathcal{G}(H^1) + \cdots + B \cap \mathcal{G}(H^k)$$

where $B = \operatorname{span}\{B_1, \cdots, B_m\}$. *Moreover, if the system* (5.7) *satisfies the controllability rank condition (i.e.,* $\mathcal{G}(\{R_e\}_G) = \mathcal{G}(G)$*), then* v^i *controls* y^i *completely.*

Proof. The proof is immediate from Lemma 5.1 and the well-known decoupling results of Nijmeijer and Schumacher [8] and Cheng [9].

6. Conclusion. We have considered a system defined on a Lie group with outputs in a coset space as described in Brockett [1]. The main results of this paper are two observability theorems, Theorems 2.2 and 2.3, that give necessary and sufficient conditions for local and global observability, respectively. Algorithm 3.1 calculates the A and B_i invariant Lie subalgebra contained in a given Lie subalgebra, which makes the condition in the above two theorems computably verifiable. Some examples are included. Finally, we have briefly discussed the input-output decoupling problem of a system on a Lie group with output in a coset space.

REFERENCES

[1] R. W. BROCKETT, *System theory on group-manifolds and coset spaces*, SIAM J. Control, 10 (1972), pp. 265-284.

[2] V. JURDJEVIC AND H. J. SUSSMANN, *Control systems on lie groups*, Differential Equations, 12 (1972), pp. 313-329.

[3] C. CHEVALLEY, *Theory of Lie Groups*, Princeton University Press, Princeton, NJ, 1946, pp. 109-111.

[4] M. HAUSNER AND J. T. SCHWARTZ, *Lie Groups; Lie Algebras*, Gordon and Breach, New York, 1968, p. 77.

[5] S. KOBAYASHI AND K. NOMIZU, *Foundations of Differential Geometry*, Vol. 1, Interscience, New York, 1963, p. 275.

[6] A. ISIDORI, *Nonlinear control systems: An introduction*, Lecture Notes in Control and Information Science 287, Springer-Verlag, Berlin, New York, 1985.

[7] A. ISIDORI, A. J. KRENER, C. GORI GIORGI, AND S. MONACO, *Nonlinear decoupling via feedback: a differential geometric approach*, IEEE Trans. Automat. Control, 26 (1981), pp. 331-345.

[8] H. NIJMEIJER AND J. M. SCHUMACHER, *Zeros at infinity for affine nonlinear control systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 566-573.

[9] D. CHENG, *Design for noninteracting decomposition of nonlinear systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 1070-1074.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 582-601, May 1990

007

OPTIMAL CONTROL OF TWO-DIMENSIONAL SYSTEMS*

MAURO BISIACCO† AND ETTORE FORNASINI‡

Abstract. Necessary and sufficient conditions for the existence and the uniqueness of the solution of the optimal control problem of discrete-time linear time invariant two-dimensional systems are determined. Given a system that satisfies these conditions, the optimal control law is obtained using an algebraic Riccati equation with coefficients in the polynomial ring $R[z]$. Since the feedback implementation of this law does not preserve the causal structure of the system, suboptimal control laws are also discussed that lead to a weakly causal two-dimensional system.

Finally, some important connections between optimal control in an l_2 setting and l_∞ stabilization are investigated.

Key words. two-dimensional systems, linear quadratic optimal control, Riccati equation, l_2 stabilizability, l_∞ stability

AMS(MOS) subject classifications. 93C55, 93D15, 49E20

1. Introduction.

Underlying a study of a two-dimensional system is often a motivation to improve its dynamical behaviour.

Recently, some authors [1]-[3] concentrated their efforts in two-dimensional state space models stabilization by means of feedback compensators. Their synthesis objective being to obtain a prescribed stable closed-loop characteristic polynomial, the approach they followed is reminiscent of the classical one-dimensional pole placement method. However, the pole placement design usually exhibits a poor control on the short-term system response, since it essentially affects the asymptotic evolution.

In this paper we take a different approach and concentrate our study on design procedures that are based on the minimization of a quadratic cost functional J. The two-dimensional system, which is the end result of this optimal design, is not merely internally stable, but satisfies additional requirements on the state and input evolutions that are summarized by the structure of J.

The class of discrete time two-dimensional systems we will consider has as a prototype the linear model described by the state updating equation [4]

$$(1.1)\quad x(h+1, k+1) = A_1 x(h, k+1) + A_2 x(h+1, k) + B_1 u(h, k+1) + B_2 u(h+1, k)$$

where $x(h, k) \in R^n$ and $u(h, k) \in R^m$ are the local state and the input value at (h, k) and A_1, A_2, B_1, B_2 are real matrices of suitable dimensions.

Assuming that the initial local states $x(i, -i)$ have been assigned, the state dynamics is completely determined by the input function $u(\cdot, \cdot)$. Our prime concern in the next section will be to pose in precise terms the optimum LQ problem for the system above. In fact, if no constraint is imposed on the structure of the infinite set of initial local states we might expect that initial conditions contribute *per se* an infinite value to the corresponding cost functional. It turns out that a satisfactory theory requires us to assume that both the space of allowed initial conditions and the space of input functions are l_2.

Within this framework, a fundamental property is the one that reduces the existence and the uniqueness of two-dimensional optimal control to a pair of rank conditions on two-dimensional polynomial matrices. This result constitutes a nontrivial extension

* Received by the editors August 3, 1988; accepted for publication (in revised form) August 5, 1989.
† Department of Mathematics and Computer Science, University of Udine, Udine, Italy.
‡ Department of Electronics and Computer Science, University of Padova, Padova, Italy.

of the corresponding one-dimensional polynomial criteria based on PBH controllability and reconstructibility matrices.

One further point of contact with the one-dimensional theory is that the optimal control can be expressed in a linear feedback form, so that the resulting closed-loop system is also a linear dynamic system. In this respect, however, a deep difference arises, since one-dimensional optimal control law preserves the original causality structure of the system, while two-dimensional causality is completely lost. This is essentially due to the fact that the optimal control value at (h, k) depends on an infinite number of local states that are not located in the past of (h, k).

The main tool for solving the problem is a special extension of the algebraic Riccati equation, whose coefficients are polynomial matrices in one variable. When the solvability conditions are satisfied, this equation admits a unique stabilizing solution, which induces a feedback matrix analytic in an open neighbourhood of the unit circle.

The analyticity property is extremely important in two respects. First, it allows us to obtain the optimal feedback law using the coefficients of a Laurent series expansion. Second, it naturally calls for an approximation procedure that provides a suboptimal control through the truncation of the above series.

A noteworthy advantage is that this suboptimal control stabilizes the closed-loop system while preserving a weakly causal two-dimensional structure, that recursively generates the feedback input values.

The l_2 approach followed in the paper is mainly motivated by the necessity of guaranteeing that the optimal control problem is a meaningful one. Usually, when dealing with the internal stability of two-dimensional systems, a more general approach is taken into account [5], since an l_∞ constraint is the only requirement imposed on the set of initial conditions. The control law we obtained by solving the l_2 optimal control problem still holds in an l_∞ setting and, interestingly enough, the l_2 stabilizability criterion, based on the rank of a two-dimensional polynomial matrix, provides a necessary and sufficient condition for l_∞ stabilizability also.

2. Problem formulation. Denote by

$$\mathscr{C}_k = \{(i, j): i+j = k\}, \qquad k = 0, 1, 2, \cdots$$

the kth *separation set* in $Z \times Z$. For purposes of future manipulation, it is useful to single out the restrictions of the input function $u(\cdot, \cdot)$ and the state evolution $x(\cdot, \cdot)$ to the separation sets and to consider such restrictions either as bilateral sequences or as Laurent formal power series in one variable. So, the restrictions of $u(\cdot, \cdot)$ and $x(\cdot, \cdot)$ to $\mathscr{C}_k$ will be denoted by the sequences

$$\mathfrak{U}_k := \{u(-i, k+i);\ i \in Z\}, \tag{2.1}$$

$$\mathfrak{X}_k := \{x(-i, k+i);\ i \in Z\}, \tag{2.2}$$

or by the series

$$\mathfrak{U}_k(\xi) := \sum_{i=\infty}^{+\infty} u(-i, k+i)\xi^i, \tag{2.3}$$

$$\mathfrak{X}_i(\xi) := \sum_{i=-\infty}^{+\infty} x(-i, k+i)\xi^i. \tag{2.4}$$

In the following, sequences (2.2) will be called *global states.*

For each initial global state $\mathfrak{X}_0$ on the separation set $\mathscr{C}_0$ and each input function $u(\cdot,\cdot)$, we introduce the quadratic *cost functional*[1]

$$J(u,\mathfrak{X}_0) := \sum_{h+k\geqq 0} [x^T(h,k)Qx(h,k)+u^T(h,k)Ru(h,k)] \tag{2.5}$$

with $R>0$ and $Q\geqq 0$. The optimal control problems we aim to solve are the following:

(i) Given $\mathfrak{X}_0$, derive conditions for the existence and the uniqueness of an input $u(\cdot,\cdot)$ that minimizes the cost J.

(ii) Whenever these conditions are satisfied, explicitly compute the optimal input and the corresponding value of J.

It is apparent from the structure of J that admissible input functions must belong to the space $l_2^{2D}(R^m)$ of R^m-valued functions $u(\cdot,\cdot)$ defined on

$$Z_+^2 := \{(h,k)\in Z\times Z: h+k\geqq 0\} = \bigcup_{k\geqq 0} \mathscr{C}_k$$

and satisfying the finite norm condition

$$\|u(\cdot,\cdot)\|_2^2 := \sum_{h+k\geqq 0} u^T(h,k)u(h,k) < \infty.$$

Furthermore, we are only interested in state dynamics $x(\cdot,\cdot)$ that belong to $l_2^{2D}(R^n)$. Although this condition is not necessary for guaranteeing the finiteness of J in case Q is singular, it fulfills the natural requirement of imposing a stable pattern on the admissible state evolutions. In fact, $x(\cdot,\cdot)\in l_2^{2D}$ implies that the associated global states $\mathfrak{X}_t$ satisfy

$$\|\mathfrak{X}_t\|_2^2 := \sum_{i=-\infty}^{+\infty} x^T(-i,t+i)x(-i,t+i) < \infty, \tag{2.6}$$

$$\sum_{t=0}^{\infty} \|\mathfrak{X}_t\|_2^2 = \|x(\cdot,\cdot)\|_2^2 \tag{2.7}$$

showing that $\|\mathfrak{X}_t\|\to 0$ as $t\to\infty$.

Just putting $t=0$ in (2.6), we argue that the allowable bilateral sequences of initial conditions must belong to $l_2(R^n)$. In this way, the optimization problem we aim to solve is completely couched in an l_2 setting.

It is well known that the infinite time least squares problem for stationary linear one-dimensional dynamical systems may be treated analytically via the algebraic Riccati equation. The questions of the existence and uniqueness of a stabilizing optimal solution, however, can be settled a priori, without explicitly solving the equation. In fact, a necessary and sufficient condition for both properties is that the polynomial matrix

$$[I-Aw \quad Bw] \tag{2.8}$$

has full rank for any w in the closed unit disk and the polynomial matrix

$$\begin{bmatrix} Q \\ I-Aw \end{bmatrix} \tag{2.9}$$

has full rank for any complex w in the unit cirle [6], [7]:

$$\gamma_1 := \{w\in C: |w|=1\}.$$

[1] Throughout the paper, ^{T} means transpose, ^{V} conjugate, and $*$ conjugate transpose.

In two-dimensional optimal control problems, the existence and uniqueness properties of a two-dimensional stabilizing control still reduce to rank conditions on polynomial matrices in two variables and the optimal control law is obtained via an algebraic Riccati equation whose coefficients are polynomial matrices in one variable. A precise statement of the main results is given by Theorem 1.

THEOREM 1. *The following facts are equivalent*:

(1) OS (*optimal solution*). *For each global state* $\mathfrak{X}_0$ *in* $l_2(R^n)$ *there exists an* l_2^{2D} *solution of the optimal control problem, i.e., there exists an input* $u(\cdot,\cdot)$ *in* $l_2^{2D}(R^m)$ *such that* $x(\cdot,\cdot)$ *belongs to* $l_2^{2D}(R^n)$ *and the corresponding value of* J *is minimized.*

(2) RC (*rank conditions*). *The two-dimensional polynomial matrix*

$$[I - A_1 z_1 - A_2 z_2 \quad B_1 z_1 + B_2 z_2] \tag{2.10}$$

has full rank on the set

$$\mathcal{M} = \{(z_1, z_2) \in C \times C : |z_1| = |z_2| \leqq 1\}$$

and the two-dimensional polynomial matrix

$$\begin{bmatrix} Q \\ I - A_1 z_1 - A_2 z_2 \end{bmatrix} \tag{2.11}$$

has full rank on the unit torus

$$\mathcal{T} = \{(z_1, z_2) \in C \times C : |z_1| = |z_2| = 1\}.$$

(3) AREz (*algebraic Riccati equation*). *The following polynomial algebraic Riccati equation*:

$$\begin{aligned} P(z) = Q &+ (A_1^T + A_2^T z^{-1}) P(z)(A_1 + A_2 z) \\ &- (A_1^T + A_2^T z^{-1}) P(z)(B_1 + B_2 z)[R + (B_1^T + B_2^T z^{-1}) P(z)(B_1 + B_2 z)]^{-1} \\ &\times (B_1^T + B_2^T z^{-1}) P(z)(A_1 + A_2 z) \end{aligned} \tag{2.12}$$

in the unknown matrix $P(z)$ *admits a unique solution in an open annulus that includes* γ_1, *with the following properties*:

(i) $P(e^{j\omega}) = P^*(e^{j\omega}) \geqq 0$, *for all* $\omega \in [0, 2\pi]$.

(ii) *The matrix*

$$K(z) := -[R + (B_1^T + B_2^T z^{-1}) P(z)(B_1 + B_2 z)]^{-1} (B_1^T + B_2^T z^{-1}) P(z)(A_1 + A_2 z) \tag{2.13}$$

is analytic in an open annulus that includes γ_1.

(iii) *The matrix*

$$\hat{\Gamma}(\omega) := (A_1 + A_2 e^{j\omega}) + (B_1 + B_2 e^{j\omega}) K(e^{j\omega}) \tag{2.14}$$

is asymptotically stable for all ω *in* $[0, 2\pi]$.

Moreover, whenever the above conditions are fulfilled, the input $u(\cdot,\cdot)$ considered in OS is uniquely determined by the stabilizing feedback law

$$u(h, k) = \sum_{i=-\infty}^{+\infty} K_i x(h + i, k - i) \tag{2.15}$$

where the matrices K_i are the coefficients of the Laurent series expansion

$$K(z) = \sum_{i=-\infty}^{+\infty} K_i z^i. \tag{2.16}$$

The implications OS→RC of the above theorem will be proved in § 3 and the implications RC→AREz→OS in § 4. In § 5 an l_∞ extension of some results will be given, while in § 6 a suboptimal solution is discussed, which seems quite attractive from the implementation point of view.

3. Necessary conditions for l_2 solvability. An obvious necessary solvability condition of the l_2 optimal control problem is that for any initial global state $\mathcal{X}_0$ in l_2, there exists some input $u(\cdot,\cdot)$ in l_2^{2D} that provides an l_2^{2D} state evolution.

Denoting by $\mathcal{V}(\mathcal{X}_0)$ the affine variety of all inputs in l_2^{2D} with this property, the above requirement is formally restated as

$$\mathcal{V}(\mathcal{X}_0) \neq \varnothing \qquad \forall \mathcal{X}_0 \in l_2. \tag{3.1}$$

It is intuitively clear, however, that the existence of input functions inducing finite values in the cost functional does not necessarily imply that the infimum of $J(\mathcal{X}_0,\cdot)$ is effectively attained for some input function in $\mathcal{V}(\mathcal{X}_0)$. So we expect that additional conditions, besides (3.1), must be fulfilled to guarantee the existence of an optimal control in $\mathcal{V}(\mathcal{X}_0)$.

The following theorem shows that in some sense it is meaningful to discuss separately the existence of l_2^{2D} state evolutions and that of optimal controls. Actually, the first problem is connected with the rank of the matrix (2.10), whereas the second depends on the rank of both (2.10) and (2.11).

The results of the theorem that provide necessary conditions for solving these problems will be supplemented by those of § 4, showing that the same conditions are also sufficient. Thus an elegant check is available for the feasibility of two-dimensional optimal control which constitutes a nontrivial extension of the results already available in the one-dimensional case.

THEOREM 2. *If $\mathcal{V}(\mathcal{X}_0) \neq \varnothing$ for all initial global states $\mathcal{X}_0 \in l_2$, the polynomial matrix* (2.10) *has full rank for all (z_1, z_2) in $\mathcal{M}$. If moreover, for each $\mathcal{X}_0 \in l_2$, there exists an input $u_{\text{opt}} \in \mathcal{V}(\mathcal{X}_0)$ such that*

$$J(\mathcal{X}_0, u) \geqq J(\mathcal{X}_0, u_{\text{opt}}) \quad \forall u \in \mathcal{V}(\mathcal{X}_0), \tag{3.2}$$

then the polynomial matrix (2.11) *has full rank for all (z_1, z_2) in $\mathcal{T}$.*

Proof. Suppose that (2.10) is not full rank at $(z_1^0, z_2^0) = (b\, e^{j\vartheta_1^0}, b\, e^{j\vartheta_2^0}) \in \mathcal{M}$.

Then there exists a nonzero vector $v \in C^n$ such that

$$v^T(I - A_1 z_1^0 - A_2 z_2^0) = 0, \qquad v^T(B_1 z_1^0 + B_2 z_2^0) = 0. \tag{3.3}$$

We now introduce the following initial global state $\mathcal{X}_0 \in l_2$:

$$\mathcal{X}_0 := \{x(i,-i) = 0 \quad \text{for } i \neq 0;\ x(0,0) = v\}$$

and suppose that there exists an input $u(\cdot,\cdot) \in l_2^{2D}$ whose corresponding state evolution $x(\cdot,\cdot)$ is in l_2^{2D}. Since $0 < b \leqq 1$, the functions u_b and x_b defined by

$$u_b(h,k) = u(h,k) b^{h+k}, \qquad h+k \geqq 0, \tag{3.4}$$

$$x_b(h,k) = x(h,k) b^{h+k}, \qquad h+k \geqq 0, \tag{3.5}$$

are in l_2^{2D} and x_b represents the state evolution determined by $\mathcal{X}_0$ and u when A_1, A_2, B_1, B_2 in (1.1) are replaced by bA_1, bA_2, bB_1, bB_2. Therefore, taking the double Fourier transforms of u_b and x_b

$$\hat{u}_b(\omega_1,\omega_2) := \sum_{h+k\geqq 0} u_b(h,k)\, e^{-ih\omega_1} e^{-ik\omega_2}, \tag{3.6}$$

$$\hat{x}_b(\omega_1,\omega_2) := \sum_{h+k\geqq 0} x_b(h,k)\, e^{-ih\omega_1} e^{-ik\omega_2}, \tag{3.7}$$

it is easy to prove that

$$v=[I-bA_1\,e^{-j\omega_1}-bA_2\,e^{-j\omega_2}]\hat{x}_b(\omega_1,\omega_2)-[bB_1\,e^{-j\omega_1}+bB_2\,e^{-j\omega_2}]\hat{u}_b(\omega_1,\omega_2) \tag{3.8}$$

holds almost everywhere in $[0,2\pi)\times[0,2\pi)$. Letting

$$a(\omega_1,\omega_2):=\begin{bmatrix}(I-bA_1^T\,e^{j\omega_1}-bA_2^T\,e^{j\omega_2})v\\(-bB_1^T\,e^{j\omega_1}-bB_2^T\,e^{j\omega_2})v\end{bmatrix}, \tag{3.9}$$

premultiplication of (3.8) by v^* gives

$$\|v\|_2^2=\left\langle a(\omega_1,\omega_2),\begin{bmatrix}\hat{u}_b(\omega_1,\omega_2)\\\hat{x}_b(\omega_1,\omega_2)\end{bmatrix}\right\rangle\leqq\|a(\omega_1,\omega_2)\|_2\cdot\left\|\begin{bmatrix}\hat{u}(\omega_1,\omega_2)\\\hat{x}(\omega_1,\omega_2)\end{bmatrix}\right\|_2$$

which in turn implies that

$$\|\hat{u}_b(\omega_1,\omega_2)\|_2^2+\|\hat{x}_b(\omega_1,\omega_2)\|_2^2\geqq\|v\|_2^4/\|a(\omega_1,\omega_2)\|_2^2. \tag{3.10}$$

Since $a(\vartheta_1^0,\vartheta_2^0)=0$, it is easy to obtain a quadratic upper bound of the following form:

$$\|a(\omega_1,\omega_2)\|_2^2\leqq M[(\omega_1-\vartheta_1^0)^2+(\omega_2-\vartheta_2^0)^2]$$

and inequality (3.10) can be replaced by

$$\|\hat{u}_b(\omega_1,\omega_2)\|_2^2+\|\hat{x}_b(\omega_1,\omega_2)\|_2^2\geqq\|v\|_2^4/M[(\omega_1-\vartheta_1^0)^2+(\omega_2-\vartheta_2^0)^2]. \tag{3.11}$$

The right-hand side of the above inequality being not summable on $[0,2\pi)\times[0,2\pi)$, we have that the same is a fortiori true for the left-hand side. This contradicts the original assumption that $u_b(\cdot,\cdot)$ and $x_b(\cdot,\cdot)$ were l_2^{2D} functions. In fact that assumption implied, by Parseval's Theorem, the summability of both $\hat{u}_b$ and $\hat{x}_b$ on $[0,2\pi)\times[0,2\pi)$.

Note that in the above proof a complex valued global state $\mathfrak{X}_0$ was considered. However the conclusion of the theorem is correct even when only real global states are allowed: we have just to analyze the dynamics that correspond to the real or imaginary part of $\mathfrak{X}_0$.

The proof of the second part of the theorem is quite long and will be omitted for sake of brevity. It may be found in [8]. □

4. A Riccati equation for two-dimensional systems. It is very well known that the algebraic Riccati equation plays a crucial role in the solution of one-dimensional optimal control problems. In this section we will derive a Riccati equation for two-dimensional systems that provides a closed-loop optimal solution to the problem of minimizing the quadratic cost functional J defined in (2.5).

As a matter of fact, the actual evaluation of the minimum cost and the explicit computation of the optimal input function lead us to study the existence of a particular solution of the Riccati equation, which we characterize in terms of positive definiteness and analyticity.

Let us first start with a preliminary analysis of the open-loop system dynamics in terms of Fourier transforms. We assume that $\mathfrak{X}_0$ and $\mathfrak{U}_t$ belong to l_2. Then, by equation (1.1) all global states $\mathfrak{X}_t$, $t=1,2,\cdots$ are in l_2 and the Fourier transforms

$$\begin{aligned}\hat{\mathfrak{U}}_1(\omega)&=\sum_{h=-\infty}^{+\infty}u(t+h,-h)\,e^{-ih\omega},\\ \hat{\mathfrak{X}}_t(\omega)&=\sum_{h=-\infty}^{+\infty}x(t+h,-h)\,e^{-ih\omega}\end{aligned} \tag{4.1}$$

have components in $L_2[0, 2\pi]$. Letting

$$\hat{A}(\omega) = A_1 + e^{j\omega}A_2, \qquad \hat{B}(\omega) = B_t + e^{j\omega}B_2,$$

equation (1.1) can be rewritten as a first-order recursive equation

$$\hat{\mathfrak{X}}_{t+1}(\omega) = \hat{A}(\omega)\hat{\mathfrak{X}}_t(\omega) + \hat{B}(\omega)\hat{\mathfrak{U}}_t(\omega), \tag{4.2}$$

whereas Parseval's and Beppo Levi's Theorems allow us to express the cost functional in the following form:

$$\begin{aligned} J &= \sum_{t=0}^{\infty} (2\pi)^{-1} \int_0^{2\pi} \mathfrak{X}_t^*(\omega) Q \hat{\mathfrak{X}}_t(\omega) + \hat{\mathfrak{U}}_t^*(\omega) R \hat{\mathfrak{U}}_t(\omega)\, d\omega \\ &= (2\pi)^{-1} \int_0^{2\pi} \sum_{t=0}^{\infty} [\hat{\mathfrak{U}}_t^*(\omega) \quad \hat{\mathfrak{X}}_t^*(\omega)] \begin{bmatrix} R & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} \hat{\mathfrak{U}}_t(\omega) \\ \hat{\mathfrak{X}}_t(\omega) \end{bmatrix} d\omega. \end{aligned} \tag{4.3}$$

Suppose that the l_2^{2D} norms of the input function $u(\cdot, \cdot)$ and of the state dynamics $x(\cdot, \cdot)$ are both finite, i.e.,

$$\begin{aligned} \|u(\cdot, \cdot)\|_2^2 &= (2\pi)^{-1} \int_0^{2\pi} \sum_{t=0}^{\infty} \hat{\mathfrak{U}}_t^*(\omega) \hat{\mathfrak{U}}_t(\omega)\, d\omega < \infty, \\ \|x(\cdot, \cdot)\|_2^2 &= (2\pi)^{-1} \int_0^{2\pi} \sum_{t=0}^{\infty} \hat{\mathfrak{X}}_t^*(\omega) \hat{\mathfrak{X}}_t(\omega)\, d\omega < \infty. \end{aligned} \tag{4.4}$$

Then, for every Hermitian matrix $\hat{P}(\omega) = \hat{P}^*(\omega)$ with elements in $L_\infty[0, 2\pi]$, we obtain the following identity:

(4.5)

$$\begin{aligned} 0 &= \hat{\mathfrak{X}}_0^*(\omega)\hat{P}(\omega)\hat{\mathfrak{X}}_0(\omega) - \sum_{t=0}^{\infty} \hat{\mathfrak{X}}_t^*(\omega)\hat{P}(\omega)\hat{\mathfrak{X}}_t(\omega) \\ &\quad + \sum_{t=0}^{\infty} [\hat{\mathfrak{U}}_t^*(\omega)\hat{B}^*(\omega) + \hat{\mathfrak{X}}_t^*(\omega)\hat{A}^*(\omega)]\hat{P}(\omega)[\hat{A}(\omega)\hat{\mathfrak{X}}_t(\omega) + \hat{B}(\omega)\hat{\mathfrak{U}}_t(\omega)] \\ &= \hat{\mathfrak{X}}_0^*(\omega)\hat{P}(\omega)\hat{\mathfrak{X}}_0(\omega) \\ &\quad + \sum_{t=0}^{\infty} [\hat{\mathfrak{U}}_t^*(\omega)\hat{\mathfrak{X}}_t^*(\omega)] \begin{bmatrix} \hat{B}^*(\omega)\hat{P}(\omega)\hat{B}(\omega) & \hat{B}^*(\omega)\hat{P}(\omega)\hat{A}(\omega) \\ \hat{A}^*(\omega)\hat{P}(\omega)\hat{B}(\omega) & \hat{A}^*(\omega)\hat{P}(\omega)\hat{A}(\omega) - \hat{P}(\omega) \end{bmatrix} \begin{bmatrix} \hat{\mathfrak{U}}_t(\omega) \\ \hat{\mathfrak{X}}_t(\omega) \end{bmatrix}. \end{aligned}$$

Integrating (4.5) between zero and 2π and adding the resulting identity to J, we obtain

$$\begin{aligned} J &= (2\pi)^{-1} \int_0^{2\pi} \hat{\mathfrak{X}}_0^*(\omega)\hat{P}(\omega)\hat{\mathfrak{X}}_0(\omega)\, d\omega \\ &\quad + (2\pi)^{-1} \int_0^{2\pi} \sum_{t=0}^{\infty} [\hat{\mathfrak{s}}_t^*(\omega) \quad \hat{\mathfrak{X}}_t^*(\omega)] \\ &\qquad \times \begin{bmatrix} R + \hat{B}^*(\omega)\hat{P}(\omega)\hat{B}(\omega) & 0 \\ 0 & \hat{E}(\omega) \end{bmatrix} \begin{bmatrix} \hat{\mathfrak{s}}_t(\omega) \\ \hat{\mathfrak{X}}_t(\omega) \end{bmatrix} \end{aligned} \tag{4.6}$$

where

$$\hat{K}(\omega) := -[R + \hat{B}^*(\omega)\hat{P}(\omega)\hat{B}(\omega)]^{-1}\hat{B}^*(\omega)\hat{P}(\omega)\hat{A}(\omega), \tag{4.7}$$

$$\hat{\mathfrak{s}}_t(\omega) := \hat{\mathfrak{U}}_t(\omega) - \hat{K}(\omega)\hat{\mathfrak{X}}_t(\omega), \tag{4.8}$$

$$\begin{aligned} \hat{E}(\omega) &= -\hat{P}(\omega) + Q + \hat{A}^*(\omega)\hat{P}(\omega)\hat{A}(\omega) - \hat{A}^*(\omega)\hat{P}(\omega)\hat{B}(\omega) \\ &\quad \times [R + \hat{B}^*(\omega)\hat{P}(\omega)\hat{B}(\omega)]^{-1}\hat{B}^*(\omega)\hat{P}(\omega)\hat{A}(\omega). \end{aligned} \tag{4.9}$$

Clearly, if we are able to choose $\hat{P}(\omega)$ in such a way that $\hat{E}(\omega)$ is zero almost everywhere in $[0, 2\pi]$, then (4.6) reduces to

$$J = (2\pi)^{-1} \int_0^{2\pi} \hat{\mathfrak{X}}_0^*(\omega)\hat{P}(\omega)\hat{\mathfrak{X}}_0(\omega)\, d\omega + (2\pi)^{-1} \int_0^{2\pi} \sum_{t=0}^{\infty} \hat{\mathfrak{s}}_t^*(\omega) \times [R + \hat{B}^*(\omega)\hat{P}(\omega)\hat{B}(\omega)]\hat{\mathfrak{s}}_t(\omega)\, d\omega$$

and the minimum value of J

$$J_{\min} = (2\pi)^{-1} \int_0^{2\pi} \hat{\mathfrak{X}}_0^*(\omega)\hat{P}(\omega)\hat{\mathfrak{X}}_0(\omega)\, d\omega \tag{4.10}$$

is attained using the closed-loop control given by

$$\hat{\mathfrak{U}}_t(\omega) = \hat{K}(\omega)\hat{\mathfrak{X}}_t(\omega). \tag{4.11}$$

The conclusion we have drawn so far depicts the situation in a way that may convince us of the intuitive reasonableness of the result. However, some caveats are in order, since the validity of the procedure depends heavily on the existence of $\hat{P}(\omega)$ and on the fact that both the input and the state dynamics given by (4.11) and (4.2) belong to l_2^{2D}.

More precisely, the solution of the optimal control problem outlined above makes sense if we are able to give a positive answer to the following questions:

(i) Is there any solution $\hat{P}(\omega) = \hat{P}^*(\omega)$ of the equation $\hat{E}(\omega) = 0$, i.e., of the ω-dependent Riccati equation (AREω)

$$\hat{P}(\omega) = Q + \hat{A}^*(\omega)\hat{P}(\omega)\hat{A}(\omega) - \hat{A}^*(\omega)\hat{P}(\omega)\hat{B}(\omega) \times [R + \hat{B}^*(\omega)\hat{P}(\omega)\hat{B}(\omega)]^{-1}\hat{B}^*(\omega)\hat{P}(\omega)\hat{A}(\omega)?$$

(ii) Among these solutions, is there any solution $\hat{P}(\omega)$ that provides, through (4.7), a feedback matrix $\hat{K}(\omega)$ mapping $L_2[0, 2\pi]$ into $L_2[0, 2\pi]$? This requirement is necessary for guaranteeing that the feedback law (4.11) (reinterpreted in the time domain) always transforms an l_2 global state $\mathfrak{X}_t$ into an l_2 input sequence $\mathfrak{U}_t$.

(iii) In particular, does (AREω) possess any (Hermitian) solution that ensures asymptotic stability of the closed-loop system, in the sense that, for any $\mathfrak{X}_0 \in \mathrm{l}_2$, the resulting global states sequence $\{\mathfrak{X}_t\}$ can be viewed as an element of l_2^{2D}? Note that this condition is needed in order to have a feedback input (4.11) that belongs to $\mathscr{V}(\mathfrak{X}_0)$.

For every fixed ω in $[0, 2\pi]$, (AREω) is the algebraic Riccati equation of a one-dimensional system over the complex field. So, if the rank conditions of Theorem 1 are fulfilled, for each ω in $[0, 2\pi]$ the equation has a unique positive semidefinite solution $\hat{P}(\omega) = \hat{P}^*(\omega)$, that makes the one-dimensional closed-loop system matrix

$$\hat{\Gamma}(\omega) = \hat{A}(\omega) + \hat{B}(\omega)\hat{K}(\omega) \tag{4.12}$$

asymptotically stable [7].

Clearly $P(\omega)$, viewed as a matrix function of ω, satisfies the first question we raised above. Actually, it provides a solution that also satisfies questions (ii) and (iii). However showing this property deserves an accurate investigation of the analytic structure of $\hat{P}(\omega)$. A first result in this direction is provided by Theorem 3, which shows that the map

$$P: \gamma_1 \to C^{n\times n}: e^{j\omega} \to \hat{P}(\omega) \tag{4.13}$$

admits an analytic extension to a suitable open annulus including γ_1, and the extension $P(z)$ satisfies (2.12).

THEOREM 3. *Assume that the matrices* (2.10) *and* (2.11) *are full rank for any* (z_1, z_2) *in* $\mathcal{M}$ *and in* $\mathcal{T}$, *respectively. Then the equation* (2.12) *admits a* (*unique*) *solution* $P(z)$ *that fulfills the following conditions*:

(i) $P(z)$ *is analytic in an open annulus that includes the unit circle* γ_1.

(ii) *For all* ω *in* $[0, 2\pi]$, $P(e^{j\omega})$ *coincides with the unique Hermitian positive-semidefinite stabilizing solution* $\hat{P}(\omega)$ *of* (AREω).

For the proof, see the Appendix.

To completely answer questions (ii) and (iii), we need to discuss certain important properties of the solution $P(z)$ obtained in Theorem 3 that shed some light on the structure of the optimal feedback law and on its stabilizing character.

(1) Because of the analytic nature of $P(z)$, there exists a Laurent series expansion

$$P(z) = \sum_{h=-\infty}^{+\infty} P_h z^h \tag{4.14}$$

that converges in an open annulus including γ_1. The coefficients P_h of (4.14) are real matrices that satisfy the conditions

$$P_h = P_{-h}^T, \qquad h = 0, 1, 2, \cdots. \tag{4.15}$$

The proof of this property depends on the following lemma.

LEMMA 1. *Let* $P(z)$ *be the solution of* (AREz) *considered in Theorem* 3. *Then*

$$P(z) = P^T(z^{-1})$$

in a suitable open annulus that includes γ_1.

Proof. For all $\omega \in [0, 2\pi]$, the matrix $P(e^{j\omega}) = \hat{P}(\omega)$ is a solution of (AREω). On the other hand, taking the transpose of both sides of (AREω) and substituting ω with $-\omega$ we check easily that $P^T(e^{-j\omega})$ is still a solution of (AREω). So $P(e^{j\omega})$ and $P^T(e^{-j\omega})$ are both Hermitian positive-semidefinite solutions of (AREω) for any ω in $[0, 2\pi]$. Because of the uniqueness of the stabilizing solution, proving that these solutions coincide reduces to show that the matrix

$$\hat{A}(\omega) - \hat{B}(\omega)[R + \hat{B}^*(\omega)P^T(^{-j\omega})\hat{B}(\omega)]^{-1}\hat{B}^*(\omega)P^T(e^{-j\omega})\hat{A}(\omega) \tag{4.16}$$

is asymptotically stable for any ω in $[0, 2\pi]$. This is again obvious, since the conjugate of (4.16) is $\hat{\Gamma}(-\omega)$, which is asymptotically stable by hypothesis.

Thus $P(z)$ and $P^T(z^{-1})$ are analytical in an open annulus that includes γ_1 and assume the same values on γ_1. By the identity principle of analytic functions, this implies $P(z) = P^T(z^{-1})$ for any z in the annulus. □

We therefore have (4.15), as an immediate consequence of the lemma. Moreover, in (4.14) the Hermiticity of $P(e^{j\omega})$ gives $P_h = P_{-h}^*$, $h = 0, 1, 2, \cdots$.

The above equalities and (4.15) imply $P_{-h}^* = P_{-h}^T$, which proves the realness of all matrices P_h.

(2) The coefficients P_h in the expansion of $P(z)$ decay exponentially as $|h|$ increases, i.e., there exist $M > 0$ and $\lambda \in (0, 1)$ such that

$$\|P_h\| < M\lambda^{|h|}, \qquad h \in Z. \tag{4.17}$$

(3) Since $P(e^{j\omega})$ is positive semidefinite for any ω in $[0, 2\pi]$,

$$R + (B_1^T + B_2^T z^{-1})P(z)(B_1 + B_2 z)$$

is invertible for every $z \in \gamma_1$ and, by a continuity argument, for every z in an open annulus that includes γ_1. Hence the matrix

$$K(z) = -[R + (B_1^T + B_2^T z^{-1})P(z)(B_1 + B_2 z)]^{-1}(B_1^T + B_2^T z^{-1})P(z)(A_1 + A_2 z)$$

extends analytically $\hat{K}(\omega)$ in the annulus and therefore admits a Laurent power series expansion

$$K(z) = \sum_{h=-\infty}^{\infty} K_h z^h. \tag{4.18}$$

Clearly, the feedback law (4.11) is well defined, since it associates an input $\hat{\mathfrak{U}}_t(\omega) \in L_2[0, 2\pi]$ to every global state $\hat{\mathfrak{X}}_t(\omega) \in L_2[0, 2\pi]$. This provides a positive answer to question (ii).

We conclude at once from these properties that the state dynamics $x(\cdot, \cdot)$ and the corresponding input function $u(\cdot, \cdot)$ are, for any initial global state $\mathfrak{X}_0$ in $\mathfrak{l}_2$, elements of $\mathfrak{l}_2^{2D}$, which is all we need to answer question (iii). Actually, the Lyapunov equation

$$\hat{V}(\omega) = I + \hat{\Gamma}^*(\omega)\hat{V}(\omega)\hat{\Gamma}(\omega) \tag{4.19}$$

admits a unique positive-definite solution, given by the sum of the following pointwise convergent series:

$$\hat{V}(\omega) = \sum_{h=0}^{\infty} \hat{\Gamma}^*(\omega)^h \hat{\Gamma}(\omega)^h.$$

Furthermore, the linearity of (4.19) and the uniqueness of its solution for every ω in $[0, 2\pi]$ imply that the matrix $\hat{V}(\omega)$ is a continuous function of ω and hence its spectral radius $\rho(\omega)$ is uniformly bounded by some positive ρ.

Combining all these properties and applying Beppo Levi's and Parseval's Theorems, we obtain

$$\|x(\cdot, \cdot)\|_2^2 = \sum_{t=0}^{\infty} \|\mathfrak{X}_t\|_2^2 = (2\pi)^{-1} \sum_{t=0}^{\infty} \int_0^{2\pi} \hat{\mathfrak{X}}_0^*(\omega)\hat{\Gamma}^*(\omega)^t \hat{\Gamma}(\omega)^t \hat{\mathfrak{X}}_0(\omega)\, d\omega$$

$$= (2\pi)^{-1} \int_0^{2\pi} \mathfrak{X}_0^*(\omega)\hat{V}(\omega)\mathfrak{X}_0(\omega)\, d\omega \leqq \rho \|\mathfrak{X}_0\|_2^2.$$

So $x(\cdot, \cdot)$ and, obviously, $u(\cdot, \cdot)$ belong to $\mathfrak{l}_2^{2D}$.

Tying together the results of Theorem 3 and its consequences, discussed at points (1)-(3), we have that the implications RC $\rightarrow$ ARE $\rightarrow$ OS in Theorem 1 are completely proved.

To conclude this section, we wish to investigate some important consequences of the time domain structure of the optimal control law

$$u(h, k) = \sum_{i=-\infty}^{+\infty} K_i x(h+i, k-i). \tag{4.20}$$

Clearly the input value at (h, k) linearly depends on the whole sequence of local states on the separation set including (h, k). So, the quarter plane causality is completely lost in the closed-loop system.

As the coefficients K_i decay to zero exponentially, it is reasonable to expect that a suboptimal control law could be achieved by truncating the infinite series (4.18) and hence by using a finite number of local states in the feedback law (4.20). This will be discussed in detail in the sequel. Here we only remark that the input (4.20) actually minimizes the cost functional, whose value is given by

$$
\begin{aligned}
J_{\min}(\mathfrak{X}_0) &= (2\pi)^{-1} \int_0^{2\pi} \hat{\mathfrak{X}}_0^*(\omega)\hat{P}(\omega)\hat{\mathfrak{X}}_0(\omega)\, d\omega \\
&= \sum_{h,k=-\infty}^{+\infty} x^T(h,-h)P_{h-k}x(k,-k) \\
&= [\cdots x^T(1,-1)x^T(0,0)x^T(-1,1)\cdots] \\
&\qquad \begin{bmatrix} \ddots & \ddots & \ddots & \\ \ddots P_0 & P_1 & P_2 \ddots & \\ \ddots P_{-1} & P_0 & P_1 \ddots & \\ \ddots P_{-2} & P_{-1} & P_0 & \\ & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} \vdots \\ x(1,-1) \\ x(0,0) \\ x(-1,1) \\ \vdots \end{bmatrix}.
\end{aligned} \tag{4.21}
$$

Remark. In case the system is autonomous (i.e., $B_1 = B_2 = 0$), the stabilizability condition (2.10) reduces to

$$\det(I - A_1 z_1 - A_2 z_2) \neq 0 \tag{4.22}$$

for $|z_1| = |z_2| \leqq 1$ or, equivalently, to the internal stability of the one-dimensional matrices $A_1 + A_2 e^{j\omega}$ for all ω in $[0, 2\pi]$. Note that the stability of $A_1 + A_2 e^{j\omega}$ for all ω is equivalent to two-dimensional internal stability [5] and hence to $\det(I - A_1 z_1 - A_2 z_2) \neq 0$ in the unit closed polydisk $\mathscr{P}_1 = \{(z_1, z_2): |z_1| \leqq 1, |z_2| \leqq 1\}$.

Under the same hypothesis, (AREω) reduces to the ω-independent Lyapunov equation for two-dimensional systems and (4.21) gives the Lyapunov function associated with its free dynamical evolution [5].

5. l_∞ stabilization. The feedback law (4.20) we obtained in the previous section, using an l_2 spaces approach, is well defined even when the initial global state $\mathfrak{X}_0$ belongs to an l_∞ space

$$\|\mathfrak{X}_0\|_\infty := \sup_{h \in Z} \|x(h,-h)\|_2 < \infty. \tag{5.1}$$

Actually, the exponential decay of the matrix sequence $\{K_i\}$ as $i \to \pm\infty$ implies that (4.20) converges for all bounded sequences of local states, which shows that $\mathfrak{U}_t$ and $\mathfrak{X}_t$ are in l_∞ for any $t \geqq 0$. Now it seems quite natural to ask whether the closed-loop asymptotic stability (4.20) realized in the l_2 case is preserved when l_∞ initial global states are allowed.

Here asymptotic stability means that $x(h,k)$ converges to zero uniformly as $h + k \to \infty$.

We first note that there exist positive constants M and λ, $0 < \lambda < 1$, such that

$$\|\hat{\Gamma}^t(\omega)\|_2 < M\lambda^t, \qquad t = 0, 1, 2, \cdots$$

for any ω in $[0, 2\pi]$ (for a proof, see Theorem 5 in the Appendix). Let us assume that all local states of $\mathfrak{X}_0$ are zero except for a single local state x. Then we have

$$\|\mathfrak{X}_t\|_\infty \leqq \|\mathfrak{X}_t\|_2 < M\lambda^t \|x\|_2. \tag{5.2}$$

In case $\mathfrak{X}_0$ is an arbitrary l_∞ sequence, any local $x(h, t-h) \in \mathfrak{X}_t$ is the superposition of $t+1$ contributions determined by the initial local states $x(h,-h)$, $x(h-1,-h+1), \cdots, x(h-t,-h+t)$. Consequently,

$$\|\mathfrak{X}_t\|_\infty = \sup_{h \in Z} \|x(h,-h+t)\|_2 < (t+1)M\lambda^t \|\mathfrak{X}_0\|_\infty \tag{5.3}$$

shows that the global states converge uniformly to zero.

Note that the above discussion implies that an l_∞ stabilizing control does exist as soon as matrix (2.10) is full rank on $\mathcal{M}$. In fact, assuming $Q = I_n$ in the cost functional, an l_2 optimal feedback law is computable through the solution of (2.12) and, as the above discussion shows, the same law provides an l_∞ stabilizing control.

It turns out that the rank condition on (2.10) is not only sufficient but also necessary for the existence of l_∞ input functions that drive any initial global state $\mathfrak{X}_0 \in l_\infty$ uniformly to zero. This is proved in the following theorem.

THEOREM 4. *Assume that the matrix* $[I - A_1 z_1 - A_2 z_2 \;\; B_1 z_1 + B_2 z_2]$ *is not full rank for some* $(z_1^0, z_2^0) \in \mathcal{M}$. *Then there exists an initial global state* $\mathfrak{X}_0 \in l_\infty$, *with* $\|\mathfrak{X}_0\|_\infty = 1$, *having the following property. For any sequence* $\{\mathfrak{U}_t\}$ *with elements in* l_∞, *the corresponding sequence of global states* $\{\mathfrak{X}_t\}$ *satisfies*

$$\|\mathfrak{X}_t\|_\infty \geqq 1, \qquad t = 1, 2, \cdots. \tag{5.4}$$

Proof. Let $(z_1^0, z_2^0) = (\rho e^{j\vartheta_1}, \rho e^{j\vartheta_2})$, $0 < \rho \leqq 1$ and define

$$\mu := e^{j(\vartheta_2 - \vartheta_1)}, \quad F := A_1 + \mu A_2, \quad G := B_1 + \mu B_2.$$

Since the one-dimensional polynomial matrix $[I - Fz \;\; G]$ is not full rank at $z = \rho e^{j\vartheta_1}$, the one-dimensional system (F, G) is not stabilizable. This implies that, modulo a change of basis in the state space, the matrices F and G have the following block structure:

$$F = \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix}, \qquad G = \begin{bmatrix} G_1 \\ 0 \end{bmatrix}$$

and the spectrum of F_{22} includes the eigenvalue $\gamma = \rho^{-1} e^{-j\vartheta_1}$.

Referring the local state space of the original two-dimensional system to the same basis and partitioning its matrices conformably with the partition of F and G

$$A_i = \begin{bmatrix} A_{11}^{(i)} & A_{12}^{(i)} \\ A_{21}^{(i)} & A_{22}^{(i)} \end{bmatrix}, \quad B_i = \begin{bmatrix} B_1^{(i)} \\ B_2^{(i)} \end{bmatrix}, \quad i = 1, 2,$$

we have

$$A_{21}^{(1)} + \mu A_{21}^{(2)} = 0, \qquad B_2^{(1)} + \mu B_2^{(2)} = 0. \tag{5.5}$$

An easy inductive argument shows that the polynomial matrices $(A_1 + A_2\xi)^r$ and $(A_1 + A_2\xi)^{r-1}(B_1 + B_2\xi)$, $r = 1, 2, \cdots$ have the following form:

$$(A_i + A_2\xi)^r = \begin{bmatrix} (*) & (*) \\ M_{21}^{(r)}(\xi)(\xi - \mu) & M_{22}^{(r)}(\xi)(\xi - \mu) + F_{22}^r \end{bmatrix}, \tag{5.6}$$

$$(A_1 + A_2\xi)^{r-1}(B_1 + B_2\xi) = \begin{bmatrix} (*) \\ N_{r-1}(\xi)(\xi - \mu) \end{bmatrix} \tag{5.7}$$

where $M_{21}^{(r)}(\xi)$, $M_{22}^{(r)}(\xi)$, $N_{r-1}(\xi)$, and $(*)$ denote polynomial matrices with elements in $C[\xi]$.

We now introduce the r-steps reachability matrix

$$\mathcal{R}_r = [(B_1 + B_2\xi) \;\; (A_1 + A_2\xi)(B_1 + B_2\xi) \cdots (A_1 + A_2\xi)^{r-1}(B_1 + B_2\xi)]. \tag{5.8}$$

Then the global state $\mathfrak{X}_r(\xi)$ that corresponds to an initial global state $\mathfrak{X}_0(\xi)$, and to inputs $\mathfrak{U}_0(\xi), \mathfrak{U}_1(\xi) \cdots \mathfrak{U}_{r-1}(\xi)$, is expressed as [9]

$$\mathfrak{X}_r(\xi) = (A_1 + A_2\xi)^r \mathfrak{X}_0(\xi) + \mathcal{R}_r(\xi) \begin{bmatrix} \mathfrak{U}_{r-1}(\xi) \\ \vdots \\ \mathfrak{U}_0(\xi) \end{bmatrix}. \tag{5.9}$$

Using (5.7), we rewrite the reachability matrix as

$$\mathscr{R}_r(\xi)=\begin{bmatrix} I & 0 \\ 0 & (\xi-\mu)I \end{bmatrix}\begin{bmatrix} (*) & (*) & (*) \\ N_0(\xi) & N_1(\xi)\cdots N_{r-1}(\xi) \end{bmatrix},$$

which shows that the forced state evolution in (5.9) has the following structure:

$$\begin{bmatrix} (*) \\ (\xi-\mu)\sum_{i=0}^{r-1} N_i(\xi)\mathfrak{U}_{r-1-i}(\xi) \end{bmatrix} := \begin{bmatrix} (*) \\ (\xi-\mu)q(\xi) \end{bmatrix}. \tag{5.10}$$

To satisfy (5.4), we introduce an initial global state

$$\mathfrak{X}_0(\xi)=\begin{bmatrix} 0 \\ v \end{bmatrix}\sum_{h=-\infty}^{+\infty}\mu^{-h}\xi^h \tag{5.11}$$

where v is a unitary eigenvector of F_{22} associated to the eigenvalue γ. Since $\mathfrak{X}_0(\xi)(\xi-\mu)=0$, it is easy to see that the corresponding free state evolution in (5.9) is

$$(A_1+A_2\xi)^r\mathfrak{X}_0(\xi)=\gamma^r\mathfrak{X}_0(\xi)+\begin{bmatrix} (*) \\ 0 \end{bmatrix}. \tag{5.12}$$

Here (*) denotes some arbitrary bilateral formal power series.

Since $\mathfrak{U}_0(\xi)$, $\mathfrak{U}_1(\xi)\cdots\mathfrak{U}_{r-1}(\xi)$ are bounded (i.e., $\|\mathfrak{U}_i\|_\infty<\infty$, $i=0,1,\cdots,r-1$), $\mathfrak{X}_r(\xi)$ is also bounded. So, we combine (5.10) and (5.12) and get the inequality

$$\begin{aligned}\|\mathfrak{X}_r(\xi)\|_\infty &= \left\|\begin{bmatrix} [*] \\ \gamma^r v\sum_k \mu^{-k}\xi^k+(\xi-\mu)q(\xi) \end{bmatrix}\right\|_\infty \\ &\geqq \|\gamma^r v\textstyle\sum_k \mu^{-k}\xi^k+(\xi-\mu)q(\xi)\|_\infty.\end{aligned} \tag{5.13}$$

One additional consequence of the boundedness of $\mathfrak{U}_i(\xi)$ is that the series $q(\xi)=\sum_k q_k\xi^k$ we introduced in (5.10) is l_∞ and therefore there exists a positive M such that $\|q_k\|<M$ for all k in Z.

Now consider in (5.13) the coefficients of the series

$$\gamma^r v\sum_{k=-\infty}^{+\infty}\mu^{-k}\xi^k+(\xi-\mu)q(\xi):=\sum_{k=-\infty}^{+\infty}g_k\xi^k.$$

It is immediate that $\mu^k g_k=v\gamma^r+\mu^k q_{k-1}-\mu^{k+1}q_k$, and summing over k yields

$$\sum_{k=0}^{N-1}g_k\mu^k=Nv\gamma^r+q_{-1}-\mu^N q_{N-1} \tag{5.14}$$

where N is an arbitrary positive integer. We therefore have

$$\begin{aligned}\|\mathfrak{X}_r(\xi)\|_\infty &\geqq \sup_{k\in Z}\|g_k\|\geqq\left(\frac{1}{N}\right)[\|g_0\|+\|g_1\|+\cdots\|g_{N-1}\|] \\ &=\left(\frac{1}{N}\right)[\|g_0\|+\|\mu g_1\|+\cdots\|\mu^{N-1}g_{N-1}\|] \\ &\geqq\left(\frac{1}{N}\right)\left\|\sum_{k=0}^{N-1}g_k\mu^k\right\|\geqq\rho^{-r}\|v\|-N^{-1}\|q_{-1}+\mu^N q_{N-1}\|\geqq\rho^{-r}-\frac{2M}{N}\end{aligned} \tag{5.15}$$

and since $N\geqq 1$ was arbitrary, $\|\mathfrak{X}_r(\xi)\|_\infty\geqq\rho^{-r}\geqq 1$. This shows that (5.4) holds independently of the choice of the inputs $\mathfrak{U}_i$ in l_∞.

By the argument used in the conclusion of Theorem 2, the statement of the theorem is correct even when only real global states are allowed. □

Remark 1. When dealing with globally reachable two-dimensional systems, i.e., systems whose reachability matrix $\mathcal{R}_n(\xi)$ is right invertible in $R(\xi)$, it has been proved in [9] that every initial global state $\mathfrak{X}_0$ can be driven to zero in a finite number of steps, irrespective of the rank condition of Theorem 4. Actually this does not involve a contradiction, since the finite-time control considered in [9] was not restricted to use only l_∞ inputs. For instance, assuming in (1.1) $A_1 = A_2 = 1$, $B_1 = -B_2 = 1$ gives a globally reachable two-dimensional system whose PBH matrix (2.10) is zero at $(z_1^0, z_2^0) = (1, 1) \in \mathcal{M}$. The global state

$$\mathfrak{X}_0(\xi) = \sum_{-\infty}^{+\infty}{}_k \xi^k$$

is controlled to zero in one step by the unbounded input

$$\mathfrak{U}_0(\xi) = -\sum_{-\infty}^{+\infty}{}_k k\xi^k.$$

Remark 2. A sufficient stabilizability condition based on the rank of (2.10) has been proved in [10] using different techniques and referring to dynamical models where the local state at (h, k) linearly depends on all local states and input values of the separation set $\mathcal{C}_{h+k-1}$.

In Kamen's paper, however, stabilizability means by definition the existence of a stabilizing state feedback, whereas the stabilizability definition considered in the present paper is essentially open loop. Actually, no a priori hypothesis has been assumed here on the way the stabilizing input functions could be generated, and the possibility of implementing the stabilizing control by a state feedback law is a theorem rather than an assumption.

The major consequence of this approach is that open-loop stabilizability, closed-loop stabilizability, and the full rank of (2.10) on $\mathcal{M}$ are equivalent properties.

6. Weakly causal suboptimal feedback. The control law (4.20), we obtained through the solution of (AREz), provides a state feedback that stabilizes (1.1) both in the l_2 and in the l_∞ settings.

Although this approach is conceptually appealing, the difficulties when no approximation is used can be very great. We already noted that, in general, the input value at (h, k) depends on the (infinitely many) local states $x(h-i, k+i)$, $i \in Z$. So, implementing (4.20) completely destroys the quarter plane causality of the original system and produces a half plane causal two-dimensional system, whose updating equation is required in principle to cope with an infinite-dimensional state vector. Moreover, to determine the solution of (2.12) is by no means a trivial task. In particular, a difficult problem that has no one-dimensional counterpart is that of obtaining the analytic structure of the feedback matrix $K(z)$ and computing the coefficients K_i that provide, in the time domain, the optimal feedback law.

We will give here two examples. The first shows how the solvability conditions based on the rank of (2.10) and (2.11) reflect into the analytic structure of $\hat{P}(\omega)$. The second gives an idea of some difficulties involved in the computation of K_is even in dimension one.

Example 1. Assume as in (1.1) $m = n = 1$ and $A_1 = B_1 = B_2 = 1$ and $A_2 = -1$. Furthermore, let $R = Q = 1$ be the weighting matrices of J.

In this case the solution of AREz can be obtained in closed form as

$$P(z) = \frac{1 \pm \sqrt{5 + 2z + 2z^{-1}}}{2(2 + z + z^{-1})}.$$

Letting $z = e^{jw}$, we obtain a negative solution and the solution

$$\hat{P}(\omega) = \frac{1}{-1+\sqrt{5+4\cos\omega}}.$$

The first cannot be taken into account, since we are looking for nonnegative solutions only; the second is positive for all ω in $[0, 2\pi]$, except at $\omega = \pi$, where $\hat{P}(\omega)$ diverges. Actually this is not surprising, because (2.10) is not full rank at $(\frac{1}{2}, -\frac{1}{2}) \in \mathcal{M}$. Hence a stabilizing optimal feedback law does not exist for some initial global state in $\mathfrak{l}_2$.

Example 2. Let us change only the sign of A_2 in the previous example. In this case the unique positive-definite solution of (AREω) is given by

$$\hat{P}(\omega) = \frac{2}{\sqrt{1+16(1+\cos\omega)^2}-(3+4\cos\omega)}$$

and the corresponding feedback matrix is

$$\hat{K}(\omega) = \frac{4(1+\cos\omega)}{1+\sqrt{1+16(1+\cos\omega)^2}}.$$

A plot of $\hat{P}(\omega)$ is given in Fig.1.

Since $\hat{P}(\omega)$ attains its minimum value at $\omega = \pi$, we have

$$J_{\min}(\mathfrak{X}_0) = (2\pi)^{-1}\int_0^{2\pi} \hat{P}(\omega)|\hat{\mathfrak{X}}_0(\omega)|^2\, d\omega \geqq \|\mathfrak{X}_0\|_2^2 \hat{P}(\pi)$$

and $J_{\min}$ can be made arbitrarily close to the lower bound if we consider initial global states whose spectral content is concentrated in a narrow neighbourhood of π.

The computation of the K_h's depends on the evaluation of the following integrals:

$$K_h = (2\pi)^{-1}\int_0^{2\pi} \frac{4(1+\cos\omega)\cos(\omega h)}{1+\sqrt{1+16(1+\cos\omega)^2}}\, d\omega.$$

Since infinitely many K_h's are different from zero, the optimal feedback law (2.15) cannot be implemented by a finite-dimensional device, and the resulting closed-loop system is a half-plane causal two-dimensional system.

To overcome the storage and computation problems, it seems natural to investigate whether, in the case (1.1) satisfying the rank condition of Theorem 3, the stabilizing feedback matrix could be constrained to have all elements in the bilateral polynomials ring $R[z, z^{-1}]$. An obvious advantage of this control law is that $u(h, k)$ would only

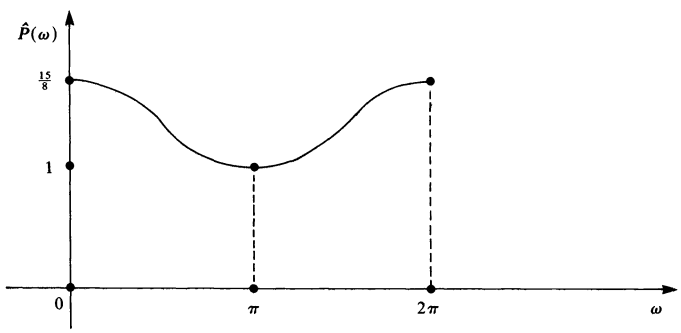


FIG. 1

depend on a finite number of local states, which makes the closed-loop system weakly causal [11].

The question above can be positively answered. Actually, we will prove that the bilateral polynomial matrix

$$K_N(z) := \sum_{i=-N}^{N} K_i z^i$$

obtained by truncation of the Laurent series (4.18) gives an (l_2 and l_∞) stabilizing state feedback, provided that N is large enough. Furthermore, when N diverges and l_2 initial states are considered, the corresponding cost functional J_N asymptotically converges to the minimum value $J_{\min}$.

To prove the first statement, recall that the coefficients K_i exponentially decay as $|i| \to \infty$. This implies that $K_N(e^{j\omega})$ and the corresponding closed-loop matrix $\Gamma_N(e^{j\omega})$ uniformly converge to $K(e^{j\omega})$ and $\Gamma(e^{j\omega})$, respectively.

Denoting by $\rho < 1$ the maximum spectral radius of $\Gamma(e^{j\omega})$ as ω varies in $[0, 2\pi]$, it will suffice to prove that for every ω the spectral radius of $\Gamma_N(e^{j\omega})$ does not exceed $(1+\rho)/2$ for large values of N. It even suffices to prove that for every ω the distance between the eigenvalues of $\Gamma_N(e^{j\omega})$ and those of $\Gamma(e^{j\omega})$ is less than $(1-\rho)/2$ for large values of N.

By the Ostrowski Theorem [12], there exists a positive real δ such that the distance between the eigenvalues of $\Gamma_N(e^{j\omega})$ and those of $\Gamma(e^{j\omega})$ is less than $(1-\rho)/2$ if

$$\|\Gamma(e^{j\omega}) - \Gamma_N(e^{j\omega})\| < \delta \tag{6.1}$$

for every ω in $[0, 2\pi]$. But the uniform convergence of $\Gamma_N(e^{j\omega})$ to $\Gamma(e^{j\omega})$ guarantees that (6.1) becomes true as N diverges.

To prove that J_N converges to $J_{\min}$ we introduce a frequency-dependent quadratic Lyapunov function that provides a very convenient integral representation of the cost functional J.

LEMMA 2. *Assume that the feedback law* $\hat{U}_t(\omega) = \hat{K}(\omega)\hat{X}_t(\omega)$ *stabilizes the system* (1.1) *in the usual* l_2 *sense and denote by* $\hat{\Gamma}(\omega)$ *the corresponding closed-loop matrix. The cost functional associated to an initial global state* $\hat{X}_0(\omega)$ *is given by*

$$J = (2\pi)^{-1} \int_0^{2\pi} \hat{X}_0^*(\omega)\hat{P}(\omega)n\hat{X}_0(\omega)\, d\omega \tag{6.2}$$

where $\hat{P}(\omega)$ *is the (unique) solution of the following Lyapunov equation*:

$$\hat{P}(\omega) = \hat{\Gamma}^*(\omega)\hat{P}(\omega)\hat{\Gamma}(\omega) + [Q + \hat{K}^*(\omega)R\hat{K}(\omega)]. \tag{6.3}$$

Proof. By Parseval's and Beppo Levi's Theorems, the cost functional can be represented as

$$J = (2\pi)^{-1} \sum_{h=0}^{\infty} \int_0^{2\pi} [\hat{X}_h^*(\omega)Q\hat{X}_h(\omega) + \hat{U}_h^*(\omega)R\hat{U}_h(\omega)]\, d\omega$$

$$= (2\pi)^{-1} \int_0^{2\pi} \hat{X}_0^*(\omega) \sum_{h=0}^{\infty} \hat{\Gamma}^*(\omega)^h [Q + \hat{K}^*(\omega)R\hat{K}(\omega)]\hat{\Gamma}(\omega)^h \hat{X}_0(\omega)\, d\omega.$$

Since $\hat{\Gamma}(\omega)$ is asymptotically stable for every ω in $[0, 2\pi]$, it is easy to check that the series

$$\sum_{h=0}^{\infty} \hat{\Gamma}^*(\omega)^h [Q + \hat{K}^*(\omega)R\hat{K}(\omega)]\hat{\Gamma}(\omega)^h$$

converges pointwise for every ω to the unique (positive-definite) solution of (6.3).

As a consequence of the above lemma, we have

$$J_N - J_{\min} = (2\pi)^{-1} \int_0^{2\pi} \hat{\mathfrak{X}}_0^*(\omega)[\hat{P}_N(\omega) - \hat{P}(\omega)]\hat{\mathfrak{X}}_0(\omega)\, d\omega. \tag{6.4}$$

Here $\hat{P}(\omega)$ is both the stabilizing solution of (AREω) and the solution of the Lyapunov equation (6.3) that includes the optimal feedback law $\hat{K}(\omega)$ and the corresponding closed-loop system matrix $\hat{\Gamma}(\omega)$. $\hat{P}_N(\omega)$ is the solution of a Lyapunov equation that includes the truncation $\hat{K}_N(\omega)$ of the optimal feedback law and the corresponding closed-loop system matrix $\hat{\Gamma}_N(\omega)$.

The matrix solution $\hat{P}(\omega)$ of (6.3), associated with the optimal feedback law, is unique and its elements $\hat{p}_{ij}(\omega)$ continuously depend on the elements of $\hat{\Gamma}(\omega)$ and $\hat{K}(\omega)$. Therefore the uniform convergence of $\hat{K}_N(\omega)$ and $\hat{\Gamma}_N(\omega)$ to $\hat{K}(\omega)$ and $\hat{\Gamma}(\omega)$ implies that $\hat{P}_N(\omega)$ uniformly converges to $\hat{P}(\omega)$. Using (6.4) we conclude that J_N converges to $J_{\min}$. □

The stabilizability condition we referred to in this paper

$$[I - A_1 z_1 - A_2 z_2 \quad B_1 z_1 + B_2 z_2] \quad \text{full rank in } \mathcal{M} \tag{6.5}$$

is weaker than the condition

$$[I - A_1 z_1 - A_2 z_2 \quad B_1 z_1 + B_2 z_2] \quad \text{full rank in } P_1, \tag{6.6}$$

which is necessary and sufficient [1] for the existence of a stabilizing state feedback law that preserves the quarter plane causality of the closed-loop system.

Clearly, in the case where (6.5) holds and (6.6) does not, losing quarter plane causality is the price we pay for achieving the closed-loop stabilization.

Although condition (6.5) is in general weaker than (6.6), if we assume $B_1 = B_2 = 0$ both conditions collapse. Actually, in this case neither causal nor noncausal feedback can stabilize the system, unless it is originally stable.

Appendix.

Proof of Theorem 3. Let $P = [p_{ij}]$ belong to $C^{n\times n}$ and introduce the map $f: C \times C^{n\times n} \to C^{n\times n}$ given by

$$\begin{aligned} f(z, P) = {} & P - Q - (A_1^T + A_2^T z^{-1})P(A_1 + A_2 z) \\ & + (A_1^T + A_2^T z^{-1})P(B_1 + B_2 z)[R + (B_1^T + B_2^T z^{-1})P(B_1 + B_2 z)]^{-1} \\ & \times (B_1^T + B_2^T z^{-1})P(A_1 + A_2 z). \end{aligned} \tag{A1}$$

We therefore have that the problem of obtaining the solutions of (2.12) reduces to that of solving, with respect to the matrix variable P, the implicit equation

$$f(z, P) = 0. \tag{A2}$$

The proof will break up into two parts. The first is devoted to a local solution of the implicit equation on the neighbourhood of an arbitrary point of the unit circle. It will be shown that, given $e^{j\bar{\omega}}$, there exists a unique analytic matrix $P_{\bar{\omega}}(\cdot)$, defined on an open disk centered in $e^{j\bar{\omega}}$, that solves (A2) and satisfies the condition

$$P_{\bar{\omega}}(e^{j\omega}) = \hat{P}(\omega).$$

The second part is concerned with the existence of a global solution of (A2). An analytic continuation $P(z)$ of the local solution will be provided on an open neighbourhood of γ_1 in such a way that the condition

$$P(e^{j\omega}) = \hat{P}(\omega)$$

holds for any $\omega \in [0, 2\pi]$.

As far as the local solution of (A2) is concerned, the definition of $\hat{P}(\omega)$ implies $f(e^{i\bar{\omega}}, \hat{P}(\bar{\omega})) = 0$, so that, to apply the Implicit Function Theorem, we must check that the Jacobian matrix of f with respect to the variables p_{ij} is nonsingular at $(e^{j\bar{\omega}}, \hat{P}(\bar{\omega}))$. Assume that the entries of P and the components of $f(z, P)$ have been lexicographically ordered, so that equation (A2) takes the following form;

$$\begin{aligned} &f_{11}(z, p_{11}, p_{12}, \cdots, p_{1n}, p_{21}, \cdots, p_{2n}, \cdots, p_{nn}) = 0, \\ &f_{12}(z, p_{11}, p_{12}, \cdots, p_{1n}, p_{21}, \cdots, p_{2n}, \cdots, p_{nn}) = 0, \\ &\cdots \\ &f_{nn}(z, p_{11}, p_{12}, \cdots, p_{1n}, p_{21}, \cdots, p_{2n}, \cdots, p_{nn}) = 0. \end{aligned} \tag{A3}$$

Letting

$$\begin{aligned} \Gamma(z, P) = (A_1 + A_2 z) - (B_1 + B_2 z)[R + (B_1^T + B_2^T z^{-1}) P (B_1 + B_2 z)]^{-1} \\ \times (B_1^T + B_2^T z^{-1}) P (A_1 + A_2 z), \end{aligned}$$

some elementary algebraic manipulations on (A1) yield the (i, j)th indexed columns of the Jacobian matrix

$$\frac{\partial f}{\partial p_{ij}} = \frac{\partial P}{\partial p_{ij}} - \Gamma^T(z^{-1}, P^T)\left(\frac{\partial P}{\partial p_{ij}}\right)\Gamma(z, P). \tag{A4}$$

In particular, if (A4) is evaluated at $(e^{j\bar{\omega}}, \hat{P}(\bar{\omega}))$, we obtain

$$\frac{\partial f}{\partial p_{ij}} = e_i e_j^T - \hat{\Gamma}^*(\bar{\omega}) e_i e_j^T \hat{\Gamma}(\bar{\omega}) \tag{A5}$$

where e_i denotes the ith column of the $n \times n$ identity matrix and $\hat{\Gamma}(\omega)$ has been defined in (4.12).

Thus, for $i, j, r, s = 1, 2, \ldots, n$, the entries of the Jacobian matrix are

$$\frac{\partial f_{rs}}{\partial p_{ij}} = \delta_{(r,s)(i,j)} - \hat{\Gamma}^V_{i,r}(\bar{\omega}) \hat{\Gamma}_{j,s}(\bar{\omega})$$

and its (r, s)th row can be expressed as

$$e_r^T \otimes e_s^T - (e_r^T \hat{\Gamma}^*(\bar{\omega})) \otimes (e_s^T \hat{\Gamma}^T(\bar{\omega})) = (e_r^T \otimes e_s^T)[I - \hat{\Gamma}^*(\bar{\omega}) \otimes \hat{\Gamma}^T(\bar{\omega})].$$

This shows that the Jacobian matrix of f with respect to P is given by

$$I - \hat{\Gamma}^*(\bar{\omega}) \otimes \hat{\Gamma}^T(\bar{\omega}) \tag{A6}$$

and is a nonsingular matrix because of the asymptotic stability of $\hat{\Gamma}(\bar{\omega})$.

Before beginning with the "global part" of the proof, we need to investigate some properties of the local solution. Since the closed-loop matrix

$$\hat{\Gamma}(\bar{\omega}) = \Gamma(e^{j\bar{\omega}}, P_{\bar{\omega}}(e^{j\bar{\omega}}))$$

is asymptotically stable, by a continuity argument $P_{\bar{\omega}}(e^{j\omega})$ is a stabilizing solution of (AREω) for any $e^{j\omega}$ in a suitable open arc $\alpha(\bar{\omega})$ of γ_1 centered in $e^{j\bar{\omega}}$.

Thus both $P_{\bar{\omega}}(e^{j\omega})$ and $\hat{P}(\omega)$ provide a stabilizing solution of (AREω) in $\alpha(\bar{\omega})$ and, by the uniqueness of the Hermitian stabilizing solution of (AREω), it suffices to prove that $P_{\bar{\omega}}(e^{j\omega})$, $e^{j\omega}$ in $\alpha(\bar{\omega})$, is an Hermitian matrix for concluding that

$$\hat{P}(\omega) = P_{\bar{\omega}}(e^{j\omega}) \quad \forall\, e^{j\omega} \in \alpha(\bar{\omega}). \tag{A7}$$

Actually, taking the conjugate transpose of the identity $f(e^{j\omega}, P_{\bar\omega}(e^{j\omega}))=0$, $e^{j\omega}$ in $\alpha(\bar\omega)$, we obtain $f(e^{j\omega}, P^*_{\bar\omega}(e^{j\omega}))=0$, $e^{j\omega}\in\alpha(\bar\omega)$. On the other hand, we have

$$P^*_{\bar\omega}(e^{j\bar\omega})=\hat P^*(\bar\omega)=\hat P(\bar\omega)=P_{\bar\omega}(e^{j\bar\omega})$$

so that the uniqueness of the solution of (A2) in a neighbourhood of $(e^{j\bar\omega}, \hat P(\bar\omega))$ implies

$$P_{\bar\omega}(e^{j\omega})=P^*_{\bar\omega}(e^{j\omega}), \qquad e^{j\omega}\in\alpha(\bar\omega)\cap D(\bar\omega)$$

where $D(\bar\omega)$ is a suitable open disk centered at $e^{j\bar\omega}$.

This last result has really been our main goal. We use it to associate with each point $e^{j\bar\omega}\in\gamma_1$ an open disk $D(\bar\omega)$, centered in $e^{j\bar\omega}$ and an analytic function $P_{\bar\omega}(z)$, defined in $D(\bar\omega)$ and satisfying $P_{\bar\omega}(e^{j\omega})=\hat P(\omega)$ on $\gamma_1\cap D(\bar\omega)$.

Extracting from the infinite open covering $\{D(\omega)\}_{\omega}\in[0,2\pi]$ a finite subcovering of γ_1 and piecing together all functions that correspond to it, we obtain a function $P(z)$ that is analytical in an open annulus including γ_1 and that fulfills the condition

(A8) $$P(e^{j\omega})=\hat P(\omega) \quad \forall\omega\in[0,2\pi].$$ □

THEOREM 5. *Let $\hat A(\cdot):[0,2\pi]\to C^{n\times n}$ be a continuous function and assume that $\hat A(\omega)$ is asymptotically stable for any ω in $[0,2\pi]$. Then there exist $M>0$ and $\lambda\in(0,1)$ such that*

$$\|\hat A^t(\omega)\|_2\leqq M\lambda^t.$$

Proof. By the continuity assumption, there exists a real $\delta>0$ such that $\hat A(\omega)(1+\delta)$ is asymptotically stable for every ω.

Then the Lyapunov equation

$$\hat P(\omega)=I+\hat A^*(\omega)\hat P(\omega)\hat A(\omega)(1+\delta)^2$$

admits a unique positive-definite solution

$$\hat P(\omega)=\sum_{t=0}^{+\infty}\hat A^*(\omega)^t\hat A(\omega)^t(1+\delta)^{2t},$$

which is continuous in $[0,2\pi]$.

Let M^2 denote the maximum spectral radius of $\hat P(\omega)$ in $[0,2\pi]$ and $\lambda:=(1+\delta)^{-1}$. Then, for every v in C^n we have

$$v^*M^2v\geqq v^*\hat P(\omega)v=\sum_{t=0}^{\infty}\|(1+\delta)^t\hat A(\omega)^tv\|_2^2\geqq\lambda^{-2t}\|\hat A(\omega)^tv\|_2^2,$$

which proves our assertion. □

REFERENCES

[1] M. BISIACCO, *State and output feedback stabilizability of* 2D *systems*, IEEE Trans. Circuits and Systems, 32 (1985), pp. 1246-1254.

[2] M. BISIACCO, E. FORNASINI, AND G. MARCHESINI, 2D *systems feedback compensation: an approach based on commutative linear transformations*, Linear Algebra Appl., 121 (1989), pp. 135-150.

[3] M. Sebek, *Polynomial approach to pole placement in* MIMO *n-D systems*, Proc. 1988 IEEE Internat. Symposium on Circuits and Systems, Vol. 1, Helsinki, June, 1988, pp. 93-96.

[4] E. FORNASINI AND G. MARCHESINI, *Doubly indexed dynamical systems: state space models and structural properties*, Math. Systems Theory, 12 (1978), pp. 59-75.

[5] ———, *Stability analysis of* 2D *systems*, IEEE Trans. Circuits and Systems, 27 (1980), pp. 1210-1217.

[6] V. KUCERA, *The discrete Riccati equation of optimal control*, Kybernetika, 8 (1972), pp. 430-447.

[7] C. E. DE SOUZA. M. R. GEVERS, AND G. C. GOODWIN, *Riccati equations in optimal filtering of nonstabilizable systems having singular state transition matrices*, Trans. Automat. Control, 31 (1986), pp. 831-838.

[8] M. BISIACCO, *Some new results in* 2D *optimal control problems*, submitted for publication.

[9] E. FORNASINI AND G. MARCHESINI, *Global properties and duality in* 2D *systems*, Systems Control Lett., 2 (1982), pp. 30-38.

[10] E. W. KAMEN, *Stabilization of linear spatially-distributed continuous time and discrete time systems*, in Multidimensional Systems Theory, Reidel, Dordrecht, 1985, pp. 101-146.

[11] J. P. GUIVER AND N. K. BOSE, *Causal and weakly causal* 2D *filters with applications in stabilization*, in Multidimensional Systems Theory, Reidel, Dordrecht, 1985, pp. 52-100.

[12] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 602-623, May 1990

008

A COMPLETE CHARACTERIZATION OF ORBIT CLOSURES OF CONTROLLABLE SINGULAR SYSTEMS UNDER RESTRICTED SYSTEM EQUIVALENCE*

D. HINRICHSEN† AND J. O'HALLORAN‡

Abstract. In this paper properties of the orbit space of controllable generalized state-space systems modulo restricted system equivalence are derived. In particular, it is shown that this space is a smooth quasiprojective variety of dimension *nm*. Then the possible degenerations of controllable systems under transformations of restricted system equivalence are characterized and it is proved that every noncontrollable system can be approximated by a family of controllable systems that belong to a single equivalence class. In the single input case, this class is uniquely determined, whereas in the multivariable case a noncontrollable system may lie in the boundary of finitely or even infinitely many equivalence classes.

Key words. singular system, restricted system equivalence, pencils, algebraic group actions, orbit spaces, orbit closures

AMS(MOS) subject classifications. 93B27, 14D20

0. Introduction.

Consider the system of differential equations

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

and the family of systems resulting from high-gain state feedback via $F_\varepsilon = [\varepsilon \quad 0]$:

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ \varepsilon & 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

(i.e., there are entries of $A + BF_\varepsilon$ that become arbitrarily large as $\varepsilon \to \infty$). Generalizing the approach of Young, Kokotovic, and Utkin in [20], we may obtain a "limit" of this family of systems by replacing each system Σ_ε with an equivalent one in such a way that a limit exists as $\varepsilon \to \infty$. "Equivalence" in this case means *restricted system equivalence* that consists of: (1) left multiplication by a nonsingular matrix L, and (2) change of coordinates in the state space via a nonsingular matrix R. Specifically, we have

$$\dot{x} = Ax + Bu \text{ is equivalent to } LR\dot{x} = LARx + LBu.$$

In the example above, we could transform each system Σ_ε by

$$L_\varepsilon = \begin{bmatrix} 1 & -1 \\ \varepsilon & 1 \end{bmatrix} \quad \text{and} \quad R_\varepsilon = \begin{bmatrix} 1/\varepsilon & 0 \\ 0 & 1/\varepsilon \end{bmatrix}$$

resulting in the family

$$\begin{bmatrix} 1/\varepsilon & -1/\varepsilon \\ 1 & 1/\varepsilon \end{bmatrix} \dot{x} = \begin{bmatrix} -1 & -1/\varepsilon \\ 1 & (\varepsilon+1)/\varepsilon \end{bmatrix} x + \begin{bmatrix} -1 \\ 1 \end{bmatrix} u.$$

Taking the limit as $\varepsilon \to \infty$, we obtain the system

$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} -1 & 0 \\ 1 & 1 \end{bmatrix} x + \begin{bmatrix} -1 \\ 1 \end{bmatrix} u.$$

* Received by the editors October 3, 1988; accepted for publication (in revised form) July 5, 1989.
† Institut für Dynamische Systeme, Universität Bremen, 2800 Bremen 33, Federal Republic of Germany.
‡ Department of Mathematical Sciences, Portland State University, Portland, Oregon 97207. The author would like to thank Universität Bremen for its support during the writing of this paper.

Note that different choices of equivalence transformations $(L_\varepsilon, R_\varepsilon)$ may result in different limits.

In general, a limit of a high-gain feedback family

$$\dot{x} = (A + BF_\varepsilon)x + Bu$$

may be given by the limit as $\varepsilon \to \infty$ of

$$L_\varepsilon R_\varepsilon \dot{x} = L_\varepsilon (A + BF_\varepsilon) R_\varepsilon x + L_\varepsilon B u$$

for some choice of equivalence transformations $(L_\varepsilon, R_\varepsilon)$. As in the example above, there is no guarantee that $\lim_{\varepsilon\to\infty} L_\varepsilon R_\varepsilon$ is nonsingular. Hence a natural setting for limits under high-gain feedback is that of *generalized state-space systems* (or semistate systems), i.e., linear systems of differential equations of the form

$$(0.1) \quad E\dot{x}(t) = Ax(t) + Bu(t), \qquad E, A \in K^{n\times n}, \quad B \in K^{n\times m} \quad (K = \mathbb{R} \text{ or } K = \mathbb{C}).$$

To secure unique solvability of (0.1) on $\mathbb{R}^+$ we always assume

$$(0.2) \qquad \det(sE - A) \not\equiv 0.$$

If E is singular, the system is called *singular*, otherwise it is called *regular*. Besides arising naturally as limits under high-gain feedback, generalized state-space systems are of interest in their own right and have received growing attention in control theory (see [16], [18], [1]–[3], [19], [15], [21]).

To provide a mathematical basis for a systematic analysis of high-gain state feedback, we must first understand limits of generalized state-space systems under transformations by restricted system equivalence. This is the basic question we address in this paper.

Besides being central to any analysis of high-gain state feedback, the determination of limits under equivalence transformations is an important part of the basic theory of generalized state-space systems. Basic system properties like controllability, observability, and stability are invariant with respect to equivalence transformations. However, every asymptotically stable system $\dot{x} = Ax$ is equivalent to a system that is arbitrarily close to an unstable system (see [10]). An analogous statement holds true for controllability. Thus, although these structural properties are invariant with respect to equivalence, they may be lost in the limit under equivalence transformations. In this paper we show that, in the case of generalized state-space systems, the limit of a controllable system under restricted system equivalence transformations is not controllable unless it is equivalent to the original system. Moreover, we characterize these uncontrollable systems that occur as limits of a given controllable system under transformations of restricted system equivalence.

This characterization of limits under equivalence transformations generalizes the results of Khadr and Martin [13] on regular systems. The natural concept of equivalence for regular state-space systems is *similarity*, which is the restriction of restricted system equivalence to the set of regular systems. Two regular state-space systems of the form

$$(0.3) \qquad \dot{x} = Ax + Bu, \qquad (A, B) \in L(n, m) := K^{n\times(n+m)} \quad (K = \mathbb{R} \text{ or } K = \mathbb{C})$$

are said to be *similar* if one can be transformed to the other via a linear change of basis in the state space K^n. In geometrical terms, this means that both systems lie in the same orbit under the *similarity action* of the general linear group $Gl_n(K)$ on $L(n, m)$:

$$(0.4) \qquad \begin{aligned} \sigma: \quad & \mathrm{Gl}_n(K) \times L(n, m) \to L(n, m), \\ & (S, (A, B)) \to S \cdot (A, B) = (SAS^{-1}, SB). \end{aligned}$$

The geometric structure of the orbit space of state-space systems under similarity has been analyzed in detail (see, for instance, [7], [8], [13], [17]). In both the case of singular systems and that of regular systems, equivalent systems may be viewed as different representations of the same physical system. Thus the set of equivalence classes may be considered as the true space of linear systems. We describe the geometry of this space, extending the theory for regular systems to singular systems.

For singular systems, the nondegeneracy condition (0.2) guarantees that the system (0.1) admits a unique distributional solution $x_{x_0,u}$ for arbitrary initial values $x(0-) = x_0 \in K^n$ and control functions $u(\cdot) \in C^\infty(\mathbb{R}_+, K^m)$. If E is singular, the solution formula for $x_{x_0,u}$ will contain the Dirac impulse and some of its derivatives for certain $x_0 \in K^n$ (see [2]). It is to this ability of the system (0.1) to produce impulsive solutions that the term "infinite-frequency behaviour" refers (see [18], [15]).

Restricted system equivalence preserves both the finite- and the infinite-frequency behaviour. It was for this reason that restricted system equivalence was introduced by Rosenbrock [16] as the natural equivalence relation on the set of generalized state-space systems. The concept of restricted system equivalence was adopted from the theory of matrix pencils (cf. [4], [16], [15]). The equivalence classes are orbits under the group action of restricted system equivalence

$$\begin{aligned} \eta: \quad & (\mathrm{Gl}_n(K) \times \mathrm{Gl}_n(K)) \times K^{n\times(2n+m)} \to K^{n\times(2n+m)}, \\ & ((L, R), (E, A, B)) \to (LER^{-1}, LAR^{-1}, LB). \end{aligned} \tag{0.5}$$

We restrict our attention to the case of singular systems of the form (0.1), i.e., without an output equation $y = Cx$. The same questions that we address in this paper can certainly be asked about singular systems with output and we could try, for example, to extend the results in [17, IV.5]. This would, however, require an additional careful analysis. Due to limitation of space, we do not consider the output question in the present paper. To provide a mathematical basis for a systematic analysis of high-gain state feedback, it suffices to pursue the above questions for systems of the form (0.1) (without an output equation).

In § 1, we develop tools and preliminary results to be used in the rest of the paper. In particular, we establish basic properties of the space of equivalence classes under left equivalence:

$$(E, A, B) \sim (LE, LA, LB), \qquad L \in \mathrm{Gl}_n(K). \tag{0.6}$$

We also discuss the concept of controllability for semistate systems and study, for any $\alpha \in K$, the map ρ_α that associates with any system (E, A, B) satisfying $\det(\alpha E - A) \neq 0$ the state-space system

$$(A_\alpha, B_\alpha) = ((\alpha E - A)^{-1}E, (\alpha E - A)^{-1}B) \in L(n, m). \tag{0.7}$$

This correspondence is obtained from a standard reparameterization of matrix pencils [4, XII, § 2] that transforms a given regular pencil (E, A) into a pencil $(\alpha E - A, E)$ in which the first matrix is nonsingular. The transformation (0.7) has been used in the context of singular systems before (see, for example, [21]).

In § 2, we study the space $S^c(n, m)/\eta$ of restricted equivalence classes of controllable systems. In particular, we show that $S^c(n, m)/\eta$ is a smooth quasiprojective variety of dimension nm.

Restricting attention to the first two components of the action (0.5), it is obvious that, in order to characterize the orbit closures of η, we must first describe the orbit closures of regular pencils under strict equivalence. In § 3 we obtain such a description by using the results of Gerstenhaber [5] about orbit closures under the conjugation

action on single matrices. As a consequence of the characterization, we see that there are only finitely many orbits contained in the orbit closure of a regular pencil.

Finally, in § 4, we obtain a complete characterization of the systems that lie in the boundary of a controllable orbit and we show that every noncontrollable system can be approximated by a family of controllable systems that all belong to a single restricted equivalence class. In the single input case ($m = 1$), this equivalence class is uniquely determined. Hasse diagrams are used to illustrate the results.

1. Preliminaries. The space of all generalized state-space equations with m input and n semistate variables is

$$S(n, m) = \{(E, A, B) \in K^{n\times(2n+m)}: \det(sE - A) \not\equiv 0\}. \tag{1.1}$$

The set $S(n, m)$ is an open dense subset of $K^{n\times(2n+m)}$ with respect to both the Zariski topology and the standard topology. In the following, we provide $S(n, m)$ with the topology induced by the standard topology on $K^{n\times(2n+m)}$. However, we will make use of the fact that $S(n, m)$ is a (Zariski) open subset of an affine variety. The space $L(n, m) = K^{n\times(n+m)}$ of systems of the form (0.3) may be embedded in the space $S(n, m)$ via the map

$$i: \quad (A, B) \to (I, A, B). \tag{1.2}$$

For $L \in \mathrm{Gl}_n(K)$, we say that (E, A, B) is left-equivalent to (LE, LA, LB). Clearly, two left-equivalent system equations $E\dot{x} = Ax + Bu$ and $LE\dot{x} = LAx + LBu$ have the same set of solutions $(x(\cdot), u(\cdot))$. Conversely, two system equations that have the same solution set are left-equivalent (see [2, Thm. 3]). Identifying in $S(n, m)$ all triples that determine the same solution set, we obtain the orbit space of the algebraic $\mathrm{Gl}_n(K)$-action:

$$\begin{aligned} \lambda: \quad & \mathrm{Gl}_n(K) \times S(n, m) \to S(n, m), \\ & (L, (E, A, B)) \to L \cdot (E, A, B) = (LE, LA, LB). \end{aligned} \tag{1.3}$$

The orbit space $\hat{S}(n, m) = S(n, m)/\lambda$ is canonically provided with the quotient topology, i.e., the finest topology for which the projection

$$\begin{aligned} \pi_\lambda: \quad & S(n, m) \to \hat{S}(n, m), \\ & (E, A, B) \to O_\lambda(E, A, B) = \{(LE, LA, LB): L \in \mathrm{Gl}_n(K)\} \end{aligned} \tag{1.4}$$

is continuous. Since $rk[E, A] = n$ by (0.2), λ is a free action of $\mathrm{Gl}_n(K)$, i.e., all the stabilizers of points in $S(n, m)$ are trivial:

$$\mathrm{Stab}_\lambda(E, A, B) = \{L \in \mathrm{Gl}_n(K): L \cdot (E, A, B) = (E, A, B)\} = \{I_n\}. \tag{1.5}$$

Because $\mathrm{Stab}_\lambda(E, A, B)$ is trivial, $O_\lambda(E, A, B)$ is isomorphic (as a variety) to $\mathrm{Gl}_n(K)$. It follows from the Closed Orbit Lemma [11, Lemma 8.3] that all the orbits $O_\lambda(E, A, B)$ are closed in $S(n, m)$.

In the following, we will embed $\hat{S}(n, m)$ into the Grassmanian $\mathrm{Grass}_n(V)$, where $V = K^{2n+m}$. The set $\mathrm{Grass}_n(V)$ of n-dimensional linear subspaces of V is endowed with the structure of a compact analytic manifold and of a projective variety. To see that $\mathrm{Grass}_n(V)$ is in fact a projective variety, consider the Plücker embedding

$$\begin{aligned} p: \quad & \mathrm{Grass}_n(V) \to \mathbb{P}(\Lambda^n V), \\ & \mathrm{span}\{v_1, \cdots, v_n\} \to v_1 \wedge \cdots \wedge v_n. \end{aligned}$$

(For more details, see [17, p. 15].) If we associate with every element (E, A, B) in $S(n, m)$ the n-dimensional linear subspace $v[E, A, B]$ of V spanned by the row vectors

of the compound matrix $[E, A, B]$, we obtain a map from the space $S(n, m)$ to $\mathrm{Grass}_n(V)$:

$$\zeta: \quad S(n, m) \to \mathrm{Grass}_n(V), \qquad (E, A, B) \to v[E, A, B]. \tag{1.6}$$

To realize the map ζ in terms of the Plücker embedding, let $I = (i_1, \cdots, i_n)$ with $1 \leqq i_1 < i_2 < \cdots < i_n \leqq 2n+m$ determine a choice of n columns of $[E, A, B]$ and let $|E, A, B|_I$ be the determinant of the $n \times n$ matrix consisting of those columns. Let $\mathcal{I}$ be the set of all subsets of $\{1, 2, \cdots, 2n+m\}$ with n elements:

$$\mathcal{I} = \{(i_1, \cdots, i_n): 1 \leqq i_1 < i_2 < \cdots < i_n \leqq 2n+m\}.$$

Then we have

$$p \circ \zeta(E, A, B) = (|E, A, B|_I)_{I \in \mathcal{I}}.$$

Thus, under the Plücker embedding of $\mathrm{Grass}_n(V)$ into $\mathbb{P}(\Lambda^n V)$, we see that ζ is a regular (polynomial) map. The induced injection

$$\hat{\zeta}: \quad \hat{S}(n, m) \to \mathrm{Grass}_n(V), \qquad O_\lambda(E, A, B) \to v[E, A, B] \tag{1.7}$$

is a homeomorphism that maps $\hat{S}(n, m)$ onto an open dense subset of the compact analytic manifold $\mathrm{Grass}_n(V)$. As such, $\hat{S}(n, m)$ is provided with the structure of a K-analytic manifold itself, and $\dim_K \hat{S}(n, m) = n(n+m)$. Since $\mathrm{Grass}_n(V)$ is a projective variety, the embedding $\hat{\zeta}$ also provides $\hat{S}(n, m)$ with the structure of a quasi-projective variety.

Of special importance in our context is the subset $S^c(n, m)$ of controllable systems (E, A, B) in $S(n, m)$. Controllability of a singular system (0.1) is usually defined via a decomposition $K^n = X_1 \oplus X_2$ of the semistate space that decomposes (0.1) into the standard form (see [4, p. 28])

$$\dot{x}_1 = A_1 x_1 + B_1 u, \qquad A_2 \dot{x}_2 = x_2 + B_2 u \tag{1.8}$$

where $\dim X_1 = \deg \det (sE - A)$, $x_1 \in X_1$, $x_2 \in X_2$, and A_2 is nilpotent. The pairs (A_1, B_1), (A_2, B_2) are uniquely determined up to similarity [6]. Following [19], we call a system (0.1) *controllable* if for any $t_1 > 0$ and all $x_0 \in X_1$, $z \in K^n$, there exists a smooth control function $u(\cdot)$ such that the solution $x(\cdot)$ of (1.1) with $x(0) = x_0$ satisfies $x(t_1) = z$. Various necessary and sufficient criteria for controllability can be found in the literature (see [16], [19], [3], [21]). In particular,

(1.9) (E, A, B) is controllable $\Leftrightarrow$ (A_1, B_1) and (A_2, B_2) are both controllable.

The following two conditions are particularly useful for our purpose.

PROPOSITION 1.1. *The system* (0.1) *is controllable if and only if it satisfies one of the following equivalent conditions*:

(a) $\mathrm{Im}\,(\alpha E - A) + \mathrm{Im}\, B = K^n$ *for all* $\alpha \in K$, *and* $\mathrm{Im}\, E + \mathrm{Im}\, B = K^n$.

(b) $\sum_{i=0}^{n-1} \mathrm{Im}\,(((\alpha E - A)^{-1} E)^i (\alpha E - A)^{-1} B) = K^n$ *for some (or all)* $\alpha \in K$ *such that* $\det(\alpha E - A) \neq 0$. □

For a proof, see [19] and [21], respectively.

It is easy to verify that controllability is invariant with respect to the group action λ. Moreover, Proposition 1.1 implies that the following subsets are open and dense in $S(n, m)$ and in $\hat{S}(n, m)$, respectively,

$$S^c(n, m) = \{(E, A, B) \in S(n, m): (E, A, B) \text{ is controllable}\},$$

$$\hat{S}^c(n, m) = \pi_\lambda(S^c(n, m)).$$

Proposition 1.1 leads us to consider the correspondence between (E, A, B) and $(A_\alpha, B_\alpha) := ((\alpha E - A)^{-1}E, (\alpha E - A)^{-1}B)$. For any $\alpha \in K$, let

$$S_\alpha(n, m) = \{(E, A, B) \in S(n, m): \det(\alpha E - A) \neq 0\}.$$

$S_\alpha(n, m)$ is an open dense subset of $S(n, m)$ for each $\alpha \in K$. Furthermore, $S_\alpha(n, m)$ is a principal open subset of $K^{n\times(2n+m)}$ (the nonzero set of a single polynomial), and so it is an affine variety (see [11, p. 10]). Because $\mathrm{Gl}_n(K)$ acts on $S_\alpha(n, m)$ with closed orbits, we conclude from a theorem of Mumford and Fogarty [14, p. 30] that the quotient $\hat{S}_\alpha(n, m)$ is an affine variety. In fact, we will see in Proposition 1.2 that $\hat{S}_\alpha(n, m)$ is isomorphic to $L(n, m)$. Because $\deg \det(sE - A) \leqq n$ for every (E, A, B) in $S(n, m)$, it follows that, for any $n+1$ distinct numbers $\alpha_1, \cdots, \alpha_{n+1}$, we have

$$S(n, m) = \bigcup_{i=1}^{n+1} S_{\alpha_i}(n, m). \tag{1.10}$$

The correspondence motivated by Proposition 1.1 results in the λ-invariant mappings

$$\rho_\alpha: \quad S_\alpha(n, m) \to L(n, m), \qquad (E, A, B) \to (A_\alpha, B_\alpha) \tag{1.11}$$

and the induced maps on the quotient sets

$$\hat{\rho}_\alpha: \quad \hat{S}_\alpha(n, m) \to L(n, m), \qquad O_\lambda(E, A, B) \to (A_\alpha, B_\alpha). \tag{1.12}$$

The following proposition, together with (1.10), establishes that, for any set of $n+1$ distinct numbers $\alpha_1, \cdots, \alpha_{n+1}$, the family $(\hat{S}_{\alpha_i}(n, m), \hat{\rho}_{\alpha_i})_{1\leqq i\leqq n+1}$ forms an atlas of analytic charts on $\hat{S}(n, m)$.

PROPOSITION 1.2. *For any $\alpha \in K$, the mapping $\rho_\alpha: S_\alpha(n, m) \to L(n, m)$ is regular, surjective, and invariant on λ-orbits, and the induced map on the quotient set*

$$\hat{\rho}_\alpha: \quad \hat{S}_\alpha(n, m) \to L(n, m)$$

is an isomorphism of varieties, which preserves controllability:

$$\hat{\rho}_\alpha(\hat{S}_\alpha^c(n, m)) = L^c(n, m). \tag{1.13}$$

Proof. The mapping ρ_α is clearly regular and λ-invariant. For any pair $(\hat{A}, \hat{B}) \in L(n, m)$, let

$$E = \hat{A}, \quad A = \alpha\hat{A} - I, \quad B = \hat{B}. \tag{1.14}$$

Then $\alpha E - A = I$ and hence $(E, A, B) \in S_\alpha(n, m)$, and we have $\rho_\alpha(E, A, B) = (\hat{A}, \hat{B})$. Therefore, the mapping ρ_α is surjective.

Now assume that $(E_i, A_i, B_i) \in S_\alpha(n, m)$, $i = 1, 2$, and that $(\alpha E_1 - A_1)^{-1}E_1 = (\alpha E_2 - A_2)^{-1}E_2$ and $(\alpha E_1 - A_1)^{-1}B_1 = (\alpha E_2 - A_2)^{-1}B_2$. If we set $L = (\alpha E_1 - A_1)(\alpha E_2 - A_2)^{-1}$, then $LB_2 = B_1$, $LE_2 = E_1$, and $L(\alpha E_2 - A_2) = \alpha E_1 - A_1$. It follows that $LA_2 = A_1$ and so $(E_1, A_1, B_1) \in O_\lambda(E_2, A_2, B_2)$. Thus the mapping $\hat{\rho}_\alpha$ is a continuous bijection. The first part of this proof shows that the mapping ρ_α has a regular right inverse, and so the inverse mapping $\hat{\rho}_\alpha^{-1}$ is regular. Finally, we see from Proposition 1.2 that the isomorphism $\hat{\rho}_\alpha$ respects controllability. □

Remark 1.3. The Liusternik–Schnirelmann category $\mathrm{cat}\, X$ of a topological space X is, by definition, the smallest cardinality of any open covering of X consisting only of subsets contractible in X (see [12] for a survey). As a consequence of the proposition, we have

$$\mathrm{cat}\, \hat{S}(n, m) \leqq n + 1. \tag{1.15}$$

2. The orbit space of semistate systems under restricted system equivalence. The group action of restricted system equivalence η described by (0.5) is an extension of the action λ to a larger transformation group. The space $S(n, m)$ is invariant with respect to η, and η induces the following algebraic $\mathrm{Gl}_n(K)$-action $\hat{\eta}$ on the quotient $\hat{S}(n, m)$:

$$\begin{aligned} \hat{\eta}: \quad & \mathrm{Gl}_n(K) \times \hat{S}(n, m) \to \hat{S}(n, m), \\ & (R, O_\lambda(E, A, B)) \to O_\lambda(ER^{-1}, AR^{-1}, B). \end{aligned} \tag{2.1}$$

The action $\hat{\eta}$ is the restriction to $\hat{S}(n, m)$ of the following algebraic action $\bar{\eta}$ on the Grassmannian:

$$\begin{aligned} \bar{\eta}: \quad & \mathrm{Gl}_n(K) \times \mathrm{Grass}_n(V) \to \mathrm{Grass}_n(V), \\ & (R, v(E, A, B)) \to v(ER^{-1}, AR^{-1}, B). \end{aligned} \tag{2.2}$$

Under the variety isomorphisms $\hat{\rho}_\alpha$ defined by (1.12), the orbits of the group action $\hat{\eta}$ are mapped onto the orbits of the similarity action (0.4). The following proposition exhibits the precise relationship between the $\hat{\eta}$-action restricted to $\hat{S}_\alpha(n, m)$

$$\hat{\eta}_\alpha = \hat{\eta} \,|\, \mathrm{Gl}_n(K) \times \hat{S}_\alpha(n, m) \tag{2.3}$$

and the similarity action σ on $L(n, m)$ defined by (0.4).

PROPOSITION 2.1. *For any $\alpha \in K$, the map $\hat{\rho}_\alpha : \hat{S}_\alpha(n, m) \to L(n, m)$ is an equivariant isomorphism of varieties with respect to the $\mathrm{Gl}_n(K)$-action $\hat{\eta}_\alpha$ on $\hat{S}_\alpha(n, m)$ and the $\mathrm{Gl}_n(K)$-action σ on $L(n, m)$.*

Proof. The fact that the map $\hat{\rho}_\alpha$ is an isomorphism was established in Proposition 1.2. It only remains to show that $\hat{\rho}_\alpha$ satisfies, for any $R \in \mathrm{Gl}_n(K)$,

$$\hat{\rho}_\alpha(\hat{\eta}_\alpha(R, O_\lambda(E, A, B))) = \sigma(R, \hat{\rho}_\alpha(O_\lambda(E, A, B))). \tag{2.4}$$

But this is a direct consequence of the following equality that is obtained directly from the definition of $\hat{\rho}_\alpha$ (see (1.3) and (1.12)):

$$\hat{\rho}_\alpha(O_\lambda(LER^{-1}, LAR^{-1}, LB)) = (RA_\alpha R^{-1}, RB_\alpha), \qquad L, R \in \mathrm{Gl}_n(K). \tag{2.5}$$

□

DEFINITION. Let G be a reductive algebraic group acting on a variety X with orbits $O(x)$, $x \in X$. A point $x \in X$ is *regular* if there is a Zariski open neighborhood U of x for which $\dim O(x) = \dim O(y)$ for all $y \in U$ (see [14, p. 10]). A point $x \in X$ is *pre-stable* if there is a G-invariant Zariski open neighborhood U of x such that the action of G on U is closed.

The following proposition characterizes the regular points of the actions η and $\hat{\eta}$.

PROPOSITION 2.2. *An element (E, A, B) of $S(n, m)$ (respectively, an element $O_\lambda(E, A, B)$ of $\hat{S}(n, m)$) has a trivial stabilizer with respect to η (respectively, $\hat{\eta}$) if and only if (E, A, B) is controllable. Furthermore, every noncontrollable system has a stabilizer of dimension greater than zero.*

Proof. If $(E, A, B) \in S_\alpha(n, m)$ is controllable and $(LER, LAR, LB) = (E, A, B)$ for some L, $R \in \mathrm{Gl}_n(K)$, then by Proposition 1.1, (A_α, B_α) is controllable, and by Proposition 2.1, $(RA_\alpha R^{-1}, RB_\alpha) = (A_\alpha, B_\alpha)$. For controllable state-space systems, the stabilizer, with respect to the similarity action, is trivial (see, for instance, [17, IV.1.4]). Therefore $R = I_n$ and $L[E, A] = [E, A]$. Since $[E, A]$ is of full rank, it follows that $L = I_n$. Thus the stabilizer of (E, A, B) (respectively, $O_\lambda(E, A, B)$) with respect to η (respectively, $\hat{\eta}$) is trivial.

Suppose (E, A, B) is not controllable. Then (A_α, B_α) is not controllable and so its controllability subspace X_1 has dimension $r < n$. Then $K^n = X_1 \oplus X_2$ and, for some $R \in \mathrm{Gl}_n(K)$, we have

$$RA_\alpha R^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \quad \text{and} \quad RB_\alpha = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}.$$

But the pair $(RA_\alpha R^{-1}, RB_\alpha)$ is stabilized by the family $(I_r \oplus aI_{n-r})_{a \in K^*}$. Therefore, the stabilizer of (A_α, B_α) has dimension greater than zero. If $S \in \mathrm{Gl}_n(K)$ stabilizes (A_α, B_α), it follows from Proposition 2.1 that $((\alpha E - A)S(\alpha E - A)^{-1}, S)$ stabilizes (E, A, B). Therefore, the stabilizers of (E, A, B) and of $O_\lambda(E, A, B)$ have dimension greater than zero. □

It follows from Proposition 2.2 that the η-orbits in $S^c(n, m)$ (respectively, the $\hat{\eta}$-orbits in $\hat{S}^c(n, m)$) are the orbits in $S(n, m)$ of maximal dimension. Thus the open set $S^c(n, m)$ (respectively, $\hat{S}^c(n, m)$) is the set of regular points relative to the action η (respectively, $\hat{\eta}$). Because the orbits in $S^c(n, m)$ (respectively, $\hat{S}^c(n, m)$) are all of the same dimension, it follows from the Closed Orbit Lemma [11, Lemma 8.3] that $\mathrm{Gl}_n(K) \times \mathrm{Gl}_n(K)$ (respectively, $\mathrm{Gl}_n(K)$) acts on $S^c(n, m)(\hat{S}(n, m))$ with closed orbits. Therefore the set of prestable points in $S(n, m)$ $(\hat{S}(n, m))$ relative to $n(\hat{\eta})$ is the set $S^c(n, m)$ $(\hat{S}^c(n, m))$. From the definition of $S^c(n, m)/\eta$ and of $\hat{S}^c(n, m)/\hat{\eta}$, it follows that they are homeomorphic.

We conclude this section by proving that the quotient $S^c(n, m)/\eta$ is a smooth quasiprojective variety. For this we need the following terminology. An increasing sequence $I = (i_1, \cdots, i_n)$ of n integers, $1 \leqq i_1 < i_2 < \cdots < i_n \leqq (n+1)m$, is called a *selection* of (n, m). Let $\mathcal{I}$ be the set of all selections of (n, m) and order $\mathcal{I}$ lexicographically. If $C(A, B) = [B, AB, \cdots, A^nB]$ is the extended controllability matrix for a controllable state-space system $(A, B) \in L^c(n, m)$ and $I \in \mathcal{I}$, we let $|C(A, B)|_I$ denote the determinant of the $n \times n$ matrix consisting of the columns of $C(A, B)$ selected by I. Let $(|C(A, B)|_I)_{I \in \mathcal{I}}$ be the sequence, ordered by the lexicographic ordering on $\mathcal{I}$, of these determinants. We say that I is a *nice selection* if, whenever the jth column of A^iB is selected by I, then the jth column of A^rB is selected by I for all r, $0 \leqq r < i$. Note that in this case i is necessarily strictly less than n.

Let $s = |\mathcal{I}| = \binom{(n+1)m}{n}$ and define a map from the σ-orbit space of $L^c(n, m)$ to projective $(s-1)$-space:

$$\psi: \quad L^c(n, m)/\sigma \to \mathbb{P}^{(s-1)}(K), \qquad O_\sigma(A, B) \to (|C(A, B)|_I)_{I \in \mathcal{I}}. \tag{2.6}$$

The map ψ is well defined because

$$|C(RAR^{-1}, RB)|_I = |RC(A, B)|_I = (\det R)|C(A, B)|_I$$

for any $R \in \mathrm{Gl}_n(K)$.

LEMMA 2.3. *The map $\psi: L^c(n, m)/\sigma \to \mathbb{P}^{(s-1)}(K)$ is injective.*

Proof. For any $(A, B) \in L^c(n, m)$, let

$$g(A, B) = \text{row span}_K\, C(A, B).$$

Then it is easy to verify that, for arbitrary $(A_i, B_i) \in L^c(n, m)$, $i = 1, 2$,

$$g(A_1, B_1) = g(A_2, B_2) \Leftrightarrow C(A_2, B_2) = RC(A_1, B_1) \text{ for some } R \in \mathrm{Gl}_n(K)$$

$$\Leftrightarrow A_2 = RA_1R^{-1} \text{ and } B_2 = RB_1 \text{ for some } R \in \mathrm{Gl}_n(K).$$

Thus the map

$$g: \quad L^c(n, m)/\sigma \to \mathrm{Grass}_n(K^{(n+1)m})$$

is injective (an algebraic embedding). Now, if $p: \mathrm{Grass}_n(K^{(n+1)m}) \to \mathbb{P}^{(s-1)}(K)$ denotes the Plücker embedding of $\mathrm{Grass}_n(K^{(n+1)m})$ (see § 1), then $\psi = p \circ g$. Hence the map ψ is injective. □

THEOREM 2.4. *The geometric quotient $\hat{S}^c(n, m)/\hat{\eta}$ is a smooth quasiprojective variety of dimension nm over K.*

Proof. We begin by showing that the quotient $\hat{S}^c(n, m)/\hat{\eta}$ is a variety. (For a general definition of variety, see [11].) In § 1 we found that, for distinct $\alpha_1, \cdots, \alpha_{n+1} \in K$, the set $\{\hat{S}_{\alpha_i}(n, m)\}_{1 \leqq i \leqq n+1}$ is an open affine cover of $\hat{S}(n, m)$. For each α_i, the map $\hat{\rho}_{\alpha_i}$ defines an isomorphism between the affine varieties $\hat{S}_{\alpha_i}(n, m)$ and $L(n, m)$ (Proposition 1.2).

Let $\{U_I\}_{I \text{ nice}}$ be the finite open affine cover of $L^c(n, m)$ described by [17, p. 52], i.e., for each nice selection I define

$$U_I = \{(A, B) \in L^c(n, m): |C(A, B)|_I \neq 0\}.$$

Let $U_i^I = \hat{\rho}_{\alpha_i}^{-1}(U_I)$. Since U_i^I is a principal open subset of the affine variety $\hat{S}_{\alpha_i}(n, m)$, it is an affine variety. Thus the set $\{U_i^I\}_{1 \leqq i \leqq n+1, I \text{ nice}}$ is a finite open affine cover of $\hat{S}^c(n, m)$. Let

$$\pi_{\hat{\eta}}: \hat{S}^c(n, m) \to \hat{S}^c(n, m)/\hat{\eta} \tag{2.7}$$

be the quotient map. The sets $V_i^I = \pi_{\hat{\eta}}(U_i^I)$, $1 \leqq i \leqq n+1$, I a nice selection, form an open cover of $\hat{S}^c(n, m)/\hat{\eta}$. Because $\mathrm{Gl}_n(K)$ acts on the affine variety U_i^I with closed orbits, the geometric quotient V_i^I is affine [14, p. 27], i.e., there is an isomorphism $\xi_i^I: V_i^I \to A_i^I$ onto an affine variety A_i^I for each index (i, I).

For each index (i, I), $1 \leqq i \leqq n+1$, $I \in \mathcal{I}$, define a map $\sigma_i^I: S^c(n, m) \to K$ by

$$\sigma_i^I(E, A, B) = |C(A_{\alpha_i}, B_{\alpha_i})|_I \det(\alpha_i E - A)^{m+1}, \tag{2.8}$$

where m is chosen so that $|C(A_{\alpha_i}, B_{\alpha_i})|_I \det(\alpha_i E - A)^m$ is a polynomial in the entries of (E, A, B) for all (i, I). For each pair of indices $((i, I), (j, J))$, $1 \leqq i, j \leqq n+1$, I, J nice selections, define a map $\sigma_{ij}^{IJ}: V_i^I \to K$ by

$$\sigma_{ij}^{IJ}(O_{\hat{\eta}}(O_\lambda(E, A, B))) = \frac{\sigma_j^J(E, A, B)}{\sigma_i^I(E, A, B)}.$$

If we let

$$A_{ij}^{IJ} = \{a \in A_i^I: \sigma_{ij}^{IJ}(\xi_i^{I^{-1}}(a)) \neq 0\},$$

then $\pi_{\hat{\eta}}^{-1}(\xi_i^{I^{-1}}(A_{ij}^{IJ})) = U_i^I \cap U_j^J = \pi_{\hat{\eta}}^{-1}(\xi_j^{J^{-1}}(A_{ji}^{JI}))$. Thus A_{ij}^{IJ} and A_{ji}^{JI} are both quotients of $U_i^I \cap U_j^J$ under the action $\hat{\eta}$. By the uniqueness of quotients, it follows that there is an isomorphism $\varphi_{ij}^{IJ}: A_{ij}^{IJ} \to A_{ji}^{JI}$ such that $\varphi_{ij}^{IJ} \circ \xi_i^I \circ \pi_{\hat{\eta}} = \xi_j^J \circ \pi_{\hat{\eta}}$ on $U_i^I \cap U_j^J$. Thus the open cover $\{V_i^I\}_{1 \leqq i \leqq n+1, I \text{ nice}}$ patches together to form a prevariety structure on $\hat{S}^c(n, m)/\hat{\eta}$. To see that $\hat{S}^c(n, m)/\hat{\eta}$ is in fact a variety, we use the fact that a prevariety is a variety if any two points lie in an affine open set [11, p. 23]. For any two elements (E_i, A_i, B_i), $i = 1, 2$, of $S^c(n, m)$, choose $\alpha \in K$ such that $(E_i, A_i, B_i) \in S_\alpha(n, m)$, $i = 1, 2$. Choose coefficients $\{a_I\}_{I \text{ nice}}$ so that the following map f is nonzero on $O_\lambda(E_i, A_i, B_i)$ for $i = 1, 2$:

$$f: \quad \hat{S}_\alpha(n, m) \to K, \qquad O_\lambda(E, A, B) \to \Sigma_{I \text{ nice}}\, a_I |C(A_\alpha, B_\alpha)|_I.$$

(To see that the map f is well defined, see equation (4.6) in § 4.) The set

$$\hat{S}_\alpha(n, m)_f = \{O_\lambda(E, A, B) \in \hat{S}_\alpha(n, m): f(O_\lambda(E, A, B)) \neq 0\}$$

is a principal open subset of $\hat{S}_\alpha(n, m)$ and so it is affine. Because $\hat{S}_\alpha(n, m)_f \subset \hat{S}^c(n, m)$ and $\mathrm{Gl}_n(K)$ acts on $\hat{S}^c(n, m)$ with closed orbits, it follows that $\pi_{\hat{\eta}}(\hat{S}_\alpha(n, m)_f)$ is an open affine set containing the two points $\pi_{\hat{\eta}}(E_i, A_i, B_i)$, $i = 1, 2$. Therefore, $\hat{S}^c(n, m)/\hat{\eta}$ is a variety.

The variety $L^c(n, m)/\sigma$ is smooth and and has dimension nm (see [7]). It follows from Proposition 2.1 that $\pi_{\hat{\eta}}(\hat{S}^c_\alpha(n, m) = \hat{S}^c_\alpha(n, m)/\hat{\eta}$ is smooth and has dimension nm. Since the open sets $\hat{S}^c_\alpha(n, m)/\hat{\eta}$ cover $\hat{S}^c(n, m)/\hat{\eta}$, the same holds true for $\hat{S}^c(n, m/\hat{\eta}$.

To show that $\hat{S}^c(n, m)/\hat{\eta}$ is quasiprojective, we embed it in projective space $\mathbb{P}^{r-1}(K)$. Let $r = \binom{(n+1)m}{n}(n+1)$, corresponding to the pairs (i, I), ordered lexicographically, where $1 \leqq i \leqq n+1$ and $I \in \mathscr{I}$. We define

$$\begin{aligned} \varphi\colon\quad & \hat{S}^c(n, m)/\hat{\eta} \to \mathbb{P}^{r-1}(K), \\ & O_{\hat{\eta}}(O_\lambda(E, A, B)) \to (\sigma_i^I(E, A, B))_{1 \leqq i \leqq n+1, I \in \mathscr{I}}, \end{aligned} \tag{2.9}$$

where the maps σ_i^I are those defined by (2.8). Clearly, φ is a polynomial map.

To see that φ is injective, we first show that φ is injective on each open set $\hat{S}^c_{\alpha_i}(n, m)/\hat{\eta}$, $1 \leqq i \leqq n+1$ (where the elements $\alpha_i \in \mathrm{K}$ are those chosen for the definition of the maps σ_i^I). It follows from Proposition 2.1 that $\hat{S}^c_{\alpha_i}(n, m)/\hat{\eta}$ and $L^c(n, m)/\sigma$ are isomorphic as varieties via

$$\hat{\hat{\rho}}_{\alpha_i}\colon\quad \hat{S}^c_{\alpha_i}(n, m)/\hat{\eta} \underset{\to}{\cong} L^c(n, m)/\sigma, \qquad O_{\hat{\eta}}(O_\lambda(E, A, B)) \to O_\sigma(A_{\alpha_i}, B_{\alpha_i}).$$

From the definitions of the maps φ, $\hat{\hat{\rho}}_{\alpha_i}$, and ψ (from Lemma 2.3), we see that $\psi \circ \hat{\hat{\rho}}_{\alpha_i} = \varphi$ on $\hat{S}^c_{\alpha_i}(n, m)/\hat{\eta}$. It follows from Lemma 2.3 that φ is injective on $\hat{S}^c_{\alpha_i}(n, m)/\hat{\eta}$.

If $\varphi(O_{\hat{\eta}}(O_\lambda(E_1, A_1, B_1))) = \varphi(O_{\hat{\eta}}(O_\lambda(E_2, A_2, B_2)))$ for some $(E_i, A_i, B_i) \in S^c(n, m)$, then choose j, $1 \leqq j \leqq n+1$, such that $\sigma_j^I(E_1, A_1, B_1) \neq 0$ for some $I \in \mathscr{I}$. It follows that the orbits $O_{\hat{\eta}}(O_\lambda(E_i, A_i, B_i))$ lie in $\hat{S}^c_{\alpha_j}(n, m)/\hat{\eta}$, $i = 1, 2$. Since φ is injective on $\hat{S}^c_{\alpha_j}(n, m)/\hat{\eta}$, we have $O_{\hat{\eta}}(O_\lambda(E_1, A_1, B_1)) = O_{\hat{\eta}}(O_\lambda(E_2, A_2, B_2))$. Thus the map φ is an embedding of $\hat{S}^c(n, m)/\hat{\eta}$ into $\mathbb{P}^{r-1}(K)$ and so the quotient $\hat{S}^c(n, m)/\hat{\eta}$ is quasiprojective. □

Remark 2.5. In [9], Helmke and Shayman show that the quotient $\hat{S}^c(n, m)/\hat{\eta}$ is compact with respect to the standard topology. This fact, along with quasi-projectivity, establishes that $\hat{S}^c(n, m)/\hat{\eta}$ is a projective variety. □

The analytic structure of the quotient $\hat{S}^c(n, m)/\hat{\eta}$ is described in the following proposition.

PROPOSITION 2.6. *$\hat{S}^c(n, m)/\hat{\eta}$ is a connected analytic manifold of dimension nm.*

Proof. The fact that $\hat{S}^c(n, m)/\hat{\eta}$ is an analytic manifold of dimension nm follows from the proof of Theorem 2.4. The fact that this manifold is connected was established in Proposition 4.4 of [6]. □

3. Orbit closures of regular pencils. We say that a system $\Sigma = (E, A, B)$ *degenerates* to a system $\bar{\Sigma} = (\bar{E}, \bar{A}, \bar{B})$ (with respect to the group action η) if $(\bar{E}, \bar{A}, \bar{B})$ is contained in the boundary of (E, A, B), i.e.,

$$(\bar{E}, \bar{A}, \bar{B}) \in \partial O_\eta(E, A, B) := \overline{O_\eta(E, A, B)} \backslash O_\eta(E, A, B)$$

where the closure is taken in the space $S(n, m)$. (Because an orbit under an algebraic group action is a constructible set [11, Lemma 8.3], the Zariski closure of an orbit is the same as closure in the standard topology.) The main goal of this paper is to characterize the orbit closures of *controllable* systems with respect to the group action η. Thus we want to determine those systems to which a given controllable system can degenerate.

For any continuous group action on a Hausdorff space, the closure of an orbit is a union of orbits. Thus we see that if Σ degenerates to $\bar{\Sigma}$, then any system in the orbit of Σ degenerates to any system in the orbit of $\bar{\Sigma}$:

$$(\bar{E}, \bar{A}, \bar{B}) \in \overline{O_\eta(E, A, B)} \Leftrightarrow O_\eta(\bar{E}, \bar{A}, \bar{B}) \subset \overline{O_\eta(E, A, B)}.$$

By restricting our attention to orbit closure within the open subsets $S_\alpha(n, m)$, we see that η-orbit closure in $S_\alpha(n, m)$ corresponds to σ-orbit closure in $L(n, m)$.

LEMMA 3.1. *If* $(\bar{E}, \bar{A}, \bar{B})$ *and* (E, A, B) *are in* $S_\alpha(n, m)$, *then*

$$(\bar{E}, \bar{A}, \bar{B}) \in \overline{O_\eta(E, A, B)} \Leftrightarrow \rho_\alpha(\bar{E}, \bar{A}, \bar{B}) \in \overline{O_\sigma(\rho_\alpha(E, A, B))}. \tag{3.1}$$

Proof. It is easy to show that η-orbit closures in $S(n, m)$ are projected by π_λ onto $\hat{\eta}$-orbit closures in $\hat{S}(n, m)$. Thus the statement (3.1) is equivalent to

$$\begin{aligned} &\pi_\lambda(\bar{E}, \bar{A}, \bar{B}) \in \overline{O_{\hat{\eta}}(\pi_\lambda(E, A, B))} \\ &\quad \Leftrightarrow \hat{\rho}_\alpha(\pi_\lambda(\bar{E}, \bar{A}, \bar{B})) \in \overline{O_\sigma(\hat{\rho}_\alpha(\pi_\lambda(E, A, B)))}. \end{aligned} \tag{3.2}$$

Since the mapping $\hat{\rho}_\alpha$ is an orbit-preserving isomorphism of varieties (Proposition 2.1), we have

$$\begin{aligned} \overline{O_\sigma(\hat{\rho}_\alpha(\pi_\lambda(E, A, B)))} &= \overline{\hat{\rho}_\alpha(O_{\hat{\eta}}(\pi_\lambda(E, A, B)))} \\ &= \hat{\rho}_\alpha[\overline{O_{\hat{\eta}}(\pi_\lambda(E, A, B))}]. \end{aligned}$$

The result (3.2) follows. □

Let $P(n)$ be the set of regular $n \times n$ matrix pencils

$$P(n) = \{(E, A) \in K^{n \times 2n} : \det(sE - A) \not\equiv 0\},$$

provided with the standard topology induced from $K^{n \times 2n}$. The open covering $\{S_\alpha(n, m)\}$ of the space $S(n, m)$ induces an open covering of the space $P(n)$: for any $\alpha \in K$, define

$$P_\alpha(n) = \{(E, A) \in P(n) : \det(\alpha E - A) \neq 0\}.$$

If $(\bar{E}, \bar{A}, \bar{B}) \in \overline{O_\eta(E, A, B)}$, then it is clear that $(\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)}$, where ε is the $\mathrm{Gl}_n(K) \times \mathrm{Gl}_n(K)$-action on the first two components:

$$\varepsilon: \quad \mathrm{Gl}_n(K) \times \mathrm{Gl}_n(K) \times P(n) \to P(n), \qquad ((L, R), (E, A)) \to (LER^{-1}, LAR^{-1}).$$

Thus any characterization of η-orbit closure in $S(n, m)$ involves a characterization of ε-orbit closure in $P(n)$. In this section we give such a characterization. The following lemma allows us to reduce the problem of ε-orbit closure to that of orbit closure under the conjugation action γ on the space $K^{n \times n}$ of single matrices, a well-known result that appears in [5].

LEMMA 3.2. *If* (E, A) *and* $(\bar{E}, \bar{A})$ *are in* $P_\alpha(n)$, *then*

$$(\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)} \Leftrightarrow \bar{A}_\alpha \in \overline{O_\gamma(A_\alpha)}. \tag{3.3}$$

Proof. If we embed $P(n)$ in $S(n, m)$ via $(E, A) \to (E, A, 0)$, then $P_\alpha(n)$ is embedded (as a closed η-invariant subset) in $S_\alpha(n, m)$. Since $(\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)}$ if and only if $(\bar{E}, \bar{A}, 0) \in \overline{O_\eta(E, A, 0)}$, it follows that (3.3) is a direct consequence of Lemma 3.1. □

LEMMA 3.3. *If* (E, A), $(\bar{E}, \bar{A}) \in P(n)$ *and* $(\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)}$, *then* $\det(s\bar{E} - \bar{A})$ *is a nonzero scalar multiple of* $\det(sE - A)$.

Proof. Suppose $\lim_{k \to \infty} L_k E R_k^{-1} = \bar{E}$ and $\lim_{k \to \infty} L_k A R_k^{-1} = \bar{A}$. Then we have

$$\begin{aligned} \det(s\bar{E} - \bar{A}) &= \lim_{k \to \infty} \det(L_k(sE - A)R_k^{-1}) \\ &= \lim_{k \to \infty} \det(L_k R_k^{-1}) \det(sE - A). \end{aligned}$$

Since $\det(\alpha\bar{E}-\bar{A})\neq 0$ for some $\alpha\in K$, it follows that $\lim_{k\to\infty}\det(L_kR_k^{-1})$ exists and

$$\lim_{k\to\infty}\det(L_kR_k^{-1})=c\neq 0.$$

The conclusion follows. □

A consequence of this lemma is

$$(E,A)\in P_\alpha(n)\Rightarrow\overline{O_\varepsilon(E,A)}\subset P_\alpha(n). \tag{3.4}$$

A necessary and sufficient condition for two regular matrix pencils to be in the same ε-orbit was first established by Weierstrass. We follow [4, XII.2] in the following presentation. In describing conjugacy classes of single matrices via the associated invariant polynomials or the elementary divisors, we consider the matrix pencil $sI-A$. To extend these ideas to strict equivalence (i.e., ε-equivalence) of pairs of matrices (E, A), we consider the homogeneous matrix pencil $sE+tA$. The $j\times j$ minors of $sE+tA$ are homogeneous polynomials of degree j in the variables s and t, with coefficients in K. Let $D_j(E, A)$ be the greatest common divisor of all of the $j\times j$ minors of $sE+tA$ (with the coefficient of the highest power of s equal to one) and let $D_0(E, A)=1$. In particular, the polynomial $D_n(E, A)=c\det(sE+tA)$, for some $c\neq 0$, is the normalized "characteristic polynomial" of the pencil (E, A). It is easy to see that

$$D_j(E,A)\,|\,D_{j+1}(E,A)\quad\text{for } 0\leqq j\leqq n-1.$$

The invariant polynomials of the matrix pencil (E, A) are defined by

$$i_k(E,A)=\frac{D_{n-k+1}(E,A)}{D_{n-k}(E,A)},\qquad 1\leqq k\leqq n. \tag{3.5}$$

The following result is an easy consequence of the criterion of Weierstrass (see [4, XII.2]).

PROPOSITION 3.4 (Weierstrass). *Two regular matrix pencils (E_1, A_1) and (E_2, A_2) are in the same ε-orbit, i.e., there exist L, $R\in \mathrm{Gl}(n, K)$ such that*

$$LE_1R=E_2\quad\text{and}\quad LA_1R=A_2,$$

if and only if both matrix pencils have the same invariant polynomials. □

The elementary divisors of a matrix pencil (E, A) are the powers of the irreducible factors obtained by decomposition of the invariant polynomials. It follows from Proposition 3.4 that the elementary divisors, up to scalar multiple (equivalently the invariant polynomials or the polynomials $D_j(E, A)$) parameterize the ε-orbits in $P(n)$.

For each irreducible homogeneous polynomial F in $K[s, t]$ and each pencil of matrices (E, A), define

$$d_{j,F}(E,A)=\text{the multiplicity of } F \text{ in } D_j(E,A),$$

$$m_{k,F}(E,A)=\text{the multiplicity of } F \text{ in } i_k(E,A).$$

Note that $d_{0,F}(E, A)=0$ and that $m_{k,F}(E, A)=d_{n-k+1,F}(E, A)-d_{n-k,F}(E, A)$, $1\leqq k\leqq n$, so that

$$\sum_{k=1}^{j} m_{k,F}(E,A)=d_{n,F}(E,A)-d_{n-j,F}(E,A),\qquad j=1,\cdots,n. \tag{3.6}$$

In particular, for each irreducible factor F of $\det(sE+tA)$, the (nonincreasing) family

$$m_F(E,A):=(m_{1,F}(E,A),\cdots,m_{n,F}(E,A))$$

is a *partition* of $d_{n,F}(E, A)$, i.e., $d_{n,F}(E,A)=\sum_{k=1}^{n} m_{k,F}(E,A)$.

Using the standard form for matrix pencils described in [4, XII.2], it can be seen that, given any homogeneous polynomial $D(s, t) = \prod_{i=0}^{h} F_i^{r_i}$ with distinct irreducible factors $F_0, \cdots, F_h$ and given a partition $p_i = (p_{i1}, \cdots, p_{in})$, $p_{i1} \geqq \cdots \geqq p_{in}$, of each r_i, $i = 0, \cdots, h$, there is a matrix pencil (E, A) in $P(n)$ with

$$m_{F_i}(E, A) = p_i \quad \text{for each } i,\ 0 \leqq i \leqq h. \tag{3.7}$$

More specifically, if $D(s, t)$ is decomposed over the field $\mathbb{C}$,

$$D(s, t) = ct^{r_0} \prod_{i=1}^{h} (s + t\lambda_i)^{r_i}, \qquad \lambda_1, \cdots, \lambda_h \text{ distinct}, \tag{3.8}$$

then a pencil satisfying (3.7) is given by

$$E = \begin{bmatrix} I_r & 0 \\ 0 & J(0, p_0) \end{bmatrix}, \qquad A = \begin{bmatrix} J(\lambda_1, p_1) & & 0 \\ & \ddots & \\ & & J(\lambda_h, p_h) \\ 0 & & & I_{n-r} \end{bmatrix}, \tag{3.9}$$

where $r = \sum_{i=1}^{h} r_i$ and $J(\lambda_i, p_i)$ is the $r_i \times r_i$ Jordan matrix with eigenvalue λ_i and block structure given by p_i.

In the sequel, we will use the following ordering on partitions of a positive integer.

DEFINITION 3.5. If $a_1 \geqq a_2 \geqq \cdots \geqq a_n \geqq 0$ and $b_1 \geqq b_2 \geqq \cdots \geqq b_n \geqq 0$ are partitions of n (i.e., $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i = n$), then we say that

$$(a_1, a_2, \cdots, a_n) \leqq (b_1, b_2, \cdots, b_n) \text{ in the } \textit{dominance order} \text{ if}$$

$$\sum_{k=1}^{j} a_k \leqq \sum_{k=1}^{j} b_k \text{ for all } j = 1, \cdots, n. \qquad \square$$

In the space $K^{n \times n}$ of single matrices, orbits under the conjugation action γ are classified by the Jordan form. It follows from the proof of Lemma 3.3 (with $E = I$ and $L_k = R_k$) that orbit closure under the conjugation action γ preserves the characteristic polynomial $\det(sI - M)$. Therefore, the orbits in the boundary of a given γ-orbit $O_\gamma(M)$ are completely described by the partitions specifying the block structure of the associated Jordan form. In fact, Gerstenhaber has shown in [5] that, for a given matrix M in $K^{n \times n}$, a matrix $\bar{M}$ is in $\overline{O_\gamma(M)}$ if and only if

$$\operatorname{rank}(\lambda I - \bar{M})^i \leqq \operatorname{rank}(\lambda I - M)^i \quad \text{for all } i \in N \text{ and all } \lambda \in \mathbb{C}. \tag{3.10}$$

In terms of the complex Jordan forms

$${}^J\bar{M} = \bigoplus_{i=1}^{h} J(\lambda_i, \bar{p}_i) \quad \text{and} \quad {}^JM = \bigoplus_{i=1}^{h} J(\lambda_i, p_i)$$

of $\bar{M}$ and M, respectively, (3.10) is equivalent to

$$\det(sI - \bar{M}) = \det(sI - M) \quad \text{and} \quad \bar{p}_i \leqq p_i, \quad i = 1, \cdots, h. \tag{3.11}$$

Since $\bar{p}_{ik}$, respectively p_{ik}, is the multiplicity of the irreducible factor $(s - \lambda_i)$ in the invariant polynomial $i_k(I, \bar{M})$, respectively, $i_k(I, M)$, we obtain from (3.6) and (3.11) that

$$\bar{M} \in \overline{O_\gamma(M)} \Leftrightarrow D_j(I, M) \mid D_j(I, \bar{M}) \quad \text{for all } j, \quad 1 \leqq j \leqq n. \tag{3.12}$$

To carry the characterization of γ-orbit closure over to the setting of matrix pencils, we consider the following degree-preserving automorphism of $K[s, t]$:

$$F(s, t) \to \tilde{F}(s, t) = F(\alpha s + t, -s), \tag{3.13}$$

where $\alpha \in K$. Given a pencil (E, A), choose α such that $(E, A) \in P_\alpha(n)$. Then this automorphism maps any $j \times j$ minor of $sE + tA$ onto the corresponding minor of $(\alpha s + t)E - sA = s(\alpha E - A) + tE$. Hence

$$\tilde{D}_j(E, A) = D_j(\alpha E - A, E). \tag{3.14}$$

A consequence of Proposition 3.4 is that the polynomials $D_j(E, A)$ are invariant on ε-orbits. In particular, we have

$$D_j(\alpha E - A, E) = D_j(I, (\alpha E - A)^{-1}E) = D_j(I, A_\alpha), \qquad (E, A) \in P_\alpha(n). \tag{3.15}$$

Remark 3.6. For a given homogeneous polynomial $D(s, t)$ of degree n, define

$$P_D = \{(E, A): \det(sE + tA) = cD(s, t) \text{ for some } 0 \neq c \in K\}.$$

Replacing $sE - A$ (respectively, $s\bar{E} - \bar{A}$) with $sE + tA$ (respectively, $s\bar{E} + t\bar{A}$) in the proof of Lemma 3.3, we see that

$$(E, A) \in P_D \quad \text{and} \quad (\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)} \Rightarrow (\bar{E}, \bar{A}) \in P_D. \tag{3.16}$$

Because the correspondence (3.13) is a degree-preserving automorphism of $K[s, t]$, we obtain from (3.14) and Proposition 3.4 that the map $(E, A) \to (\alpha E - A, E)$ induces a bijective correspondence between the ε-orbits in P_D and the ε-orbits in $P_{\tilde{D}}$.

The following result characterizes the ε-orbit closures in the pencil space $P(n)$.

THEOREM 3.7. *Let* $(\bar{E}, \bar{A})$, $(E, A) \in P(n)$. $(\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)}$ *if and only if one of the following equivalent conditions is satisfied*:

(i) $D_j(E, A) \mid D_j(\bar{E}, \bar{A})$ *for all* j, $1 \leqq j \leqq n$.

(ii) $m_F(\bar{E}, \bar{A}) \leqq m_F(E, A)$ *for all irreducible factors* F *of* $\det(sE + tA)$.

Proof. It follows from (3.16) that we need only consider the set P_D for a fixed homogeneous polynomial D. Choose $\alpha \in K$ so that $D(\alpha, -1) \neq 0$, i.e., $P_D \subset P_\alpha(n)$. By Lemma 3.2, we have

$$(\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)} \Leftrightarrow \bar{A}_\alpha \in \overline{O_\gamma(A_\alpha)}.$$

From (3.12), we have

$$\bar{A}_\alpha \in \overline{O_\gamma(A_\alpha)} \Leftrightarrow D_j(I, A_\alpha) \mid D_j(I, \bar{A}_\alpha) \quad \text{for all } j, \quad 1 \leqq j \leqq n.$$

From (3.14), (3.15), and the fact that the correspondence (3.13) is a degree-preserving automorphism of $K[s, t]$, it follows that

$$D_j(I, A_\alpha) \mid D_j(I, \bar{A}_\alpha) \Leftrightarrow D_j(E, A) \mid D_j(\bar{E}, \bar{A}).$$

Therefore, $(\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)}$ if and only if condition (i) is satisfied. To see that conditions (i) and (ii) are equivalent, we note that condition (i) is equivalent to

$$\begin{aligned} &d_{j,F}(E, A) \leqq d_{j,F}(\bar{E}, \bar{A}) \quad \text{for all } j, \quad 1 \leqq j \leqq n, \\ &\text{and for all irreducible factors } F \text{ of } \det(sE + tA), \end{aligned} \tag{3.17}$$

hence by equation (3.6), to condition (ii). □

In the case $E = \bar{E} = I$, Theorem 3.7 coincides with Gerstenhaber's Theorem in [5].

The following is a direct consequence of the theorem.

COROLLARY 3.8. *Let* $(E, A) \in P(n)$. *Then*

(i) $\overline{O_\varepsilon(E, A)}$ *contains only finitely many orbits,*

(ii) $O_\varepsilon(E, A)$ *is closed in* $P(n)$ *if and only if* $m_{k,F}(E, A) \leqq 1$ *for all* k, $1 \leqq k \leqq n$, *and for all irreducible factors* F *of* $\det(sE + tA)$.

Remark 3.9. Using the standard form (3.9), it is clear from part (ii) of the corollary that every closed orbit contains a pencil of the form

$$\left[\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} S & 0 \\ 0 & I_{n-r} \end{bmatrix}\right] \tag{3.18}$$

where S is semisimple (diagonalizable over $\mathbb{C}$).

To better understand the consequences of Theorem 3.7, we introduce the concept of Hasse diagrams for orbit closures. A Hasse diagram is a directed graph whose vertices represent orbits and whose directed edges represent orbit closures. More precisely, the orbit at the terminal node of an arrow is in the closure of the orbit at the initial node of the arrow. Of course the diagram is transitive. It follows from (3.16) that there are an infinite number of disjoint components in the Hasse diagram of ε-orbits, each component consisting of ε-orbits in P_D for a fixed homogeneous polynomial $D(s, t)$ of degree n in $K[s, t]$. Because ε-orbits are classified by partitions of the multiplicities of the irreducible factors of D, P_D contains only finitely many orbits. For any positive integer m, $m \leqq n$, there is a unique largest partition of m (with respect to the dominance order) $p^* = (m, 0, \cdots, 0) \in \mathbb{N}^n$ and a unique smallest one $p_* = (1, \cdots, 1, 0, \cdot, 0) \in \mathbb{N}^n$. It follows that there are unique ε-orbits O^* and O_* in P_D such that $P_D = \overline{O^*}$ and every ε-orbit in P_D has O_* in its closure. In particular, the Hasse diagram for P_D is connected and has unique maximal and minimal vertices. These smallest orbits O_* (of minimal dimension within P_D) are the closed ε-orbits, which we described in Remark 3.9. From Theorem 3.7(ii), we see that the maximal orbit O^* in P_D must be the set of pencils (E, A) in P_D that satisfy the following:

$$m_{k,F}(E, A) = 0 \text{ for } k > 1 \text{ and for all irreducible factors } F \text{ of } \det(sE + tA). \tag{3.19}$$

Using the standard form (3.9), we see that the pencils which satisfy (3.19) are those that are equivalent to pencils of the form

$$\left[\begin{bmatrix} I_r & 0 \\ 0 & N \end{bmatrix}, \begin{bmatrix} J & 0 \\ 0 & I_{n-r} \end{bmatrix}\right] \tag{3.20}$$

where N is a nilpotent matrix with one Jordan block and J has one Jordan block for each eigenvalue. We illustrate these ideas with the following example.

Example 3.10. In the case $D = s^2t^2$, the set P_D consists of pencils of 4×4 matrices (E, A) such that $\det(sE + tA) = cs^2t^2$ for some $0 \neq c \in K$. Figure 1 is the Hasse diagram for P_D, where each vertex is given by a representative from an ε-orbit. We denote the 2×2 Jordan block with eigenvalue 0 by J_2 and the 2×2 identity matrix by I_2.

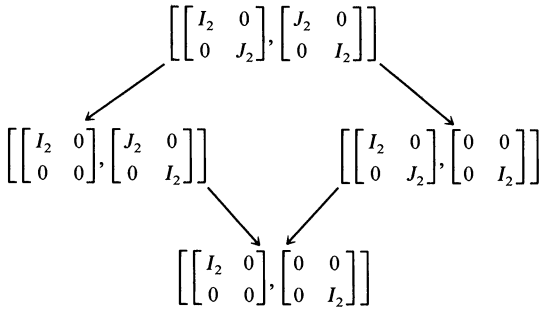


FIG. 1.

4. Orbit closures of controllable semistate systems. In this section, we describe orbit closures of controllable systems (E, A, B) under the restricted equivalence action η given by

$$(L, R) \cdot (E, A, B) = (LER^{-1}, LAR^{-1}, LB).$$

We will denote the triple (LER^{-1}, LAR^{-1}, LB) by $L(E, A, B)R^{-1}$.

Analogous to the case of regular pencils that was studied in § 3, we can restrict our attention to $S_\alpha(n, m)$; it follows from (3.4) that

$$(E, A, B) \in S_\alpha(n, m) \Rightarrow \overline{O_\eta(E, A, B)} \subset S_\alpha(n, m). \tag{4.1}$$

Thus we can reduce the problem of η-orbit closures in $S(n, m)$ to that of η-orbit closures in $S_\alpha(n, m)$, $\alpha \in K$. From Lemma 3.1, we see that η-orbit closures in $S_\alpha(n, m)$ correspond to σ-orbit closures in $L(n, m)$ via the map ρ_α defined by (1.11).

LEMMA 4.1. *Suppose that* $(\bar{E}, \bar{A}, \bar{B})$ *and* (E, A, B) *are in* $S(n, m)$ *and that* (L_k) *and* (R_k) *are sequences in* $\mathrm{Gl}_n(K)$ *such that*

$$\lim_{k\to\infty} L_k(E, A, B)R_k^{-1} = (\bar{E}, \bar{A}, \bar{B}). \tag{4.2}$$

If (E, A, B) *is controllable, then*

$$Y = \lim_{k\to\infty} L_k \quad \textit{and} \quad X = \lim_{k\to\infty} R_k \tag{4.3}$$

exist in $K^{n\times n}$ *and*

$$YE = \bar{E}X, \quad YA = \bar{A}X, \quad \bar{B} = YB. \tag{4.4}$$

Proof. Choose any $\alpha \in K$ such that $\det(\alpha E - A) \neq 0$. Then, by (4.1), $\det(\alpha\bar{E} - \bar{A}) \neq 0$. The α-controllability matrix of $(E, A, B) \in S_\alpha(n, m)$ is, by definition, the controllability matrix of $\rho_\alpha(E, A, B) = (A_\alpha, B_\alpha)$:

$$C_\alpha(E, A, B) = [B_\alpha, A_\alpha B_\alpha, \cdots, A_\alpha^{n-1}B_\alpha]. \tag{4.5}$$

By direct computation, we obtain

$$C_\alpha(L_k(E, A, B)R_k^{-1}) = R_k C_\alpha(E, A, B), \tag{4.6}$$

and from (4.2) it follows that

$$\lim_{k\to\infty} R_k C_\alpha(E, A, B) = C_\alpha(\bar{E}, \bar{A}, \bar{B}) \tag{4.7}$$

because the mapping $C_\alpha : (E, A, B) \to C_\alpha(E, A, B)$ is continuous on $S_\alpha(n, m)$. Since (E, A, B) is controllable, it follows from Proposition 1.1 that $C_\alpha(E, A, B)$ has rank n and hence $X = \lim_{k\to\infty} R_k$ exists. By the hypothesis (4.2), we have

$$\lim_{k\to\infty} L_k(\alpha E - A)R_k^{-1} = \alpha\bar{E} - \bar{A}, \tag{4.8}$$

hence

$$\lim_{k\to\infty} L_k(\alpha E - A) = (\alpha\bar{E} - \bar{A})X. \tag{4.9}$$

Since $\det(\alpha E - A) \neq 0$, it follows that $Y = \lim_{k\to\infty} L_k$ exists and we have

$$Y(\alpha E - A) = (\alpha\bar{E} - \bar{A})X, \qquad YB = \bar{B}. \tag{4.10}$$

Since (4.10) holds for every α such that $\det(\alpha\bar{E} - \bar{A}) \neq 0$, the equalities (4.4) follow. □

Remark 4.2. Let $(\bar{E}, \bar{A}, \bar{B}), (E, A, B) \in S_\alpha(n, m)$ and suppose that they satisfy (4.4). Then

$$Y = (\alpha\bar{E} - \bar{A})X(\alpha E - A)^{-1} \tag{4.11}$$

and we can rewrite (4.4) in terms of A_α and B_α:

$$\begin{aligned} &XA_\alpha = \bar{A}_\alpha X, \qquad X(\alpha E - A)^{-1}A = (\alpha\bar{E} - \bar{A})^{-1}\bar{A}X, \\ &XB_\alpha = \bar{B}_\alpha. \end{aligned} \tag{4.12}$$

Conversely, the existence of $X \in K^{n\times n}$ satisfying (4.12) implies the equalities (4.4) where Y is defined by (4.11). Using the equalities (4.12) and induction, it is straightforward to show that

$$XC_\alpha(E, A, B) = C_\alpha(\bar{E}, \bar{A}, \bar{B}). \tag{4.13}$$

If the system (E, A, B) is controllable, then $C_\alpha(E, A, B)$ has full rank and it follows that X is unique. It then follows from (4.11) that Y is unique.

We are now in a position to describe the η-orbit closures of controllable systems in $S(n, m)$.

THEOREM 4.3. *Let* $(\bar{E}, \bar{A}, \bar{B}), (E, A, B) \in S(n, m)$ *and suppose that* (E, A, B) *is controllable. Then* $(\bar{E}, \bar{A}, \bar{B}) \in \overline{O_\eta(E, A, B)}$ *if and only if the following two conditions are satisfied*:

(i) *There exists a unique pair of matrices* $X, Y \in K^{n\times n}$ *such that* $YE = \bar{E}X$, $YA = \bar{A}X$, *and* $\bar{B} = YB$.

(ii) $D_j(E, A)$ *divides* $D_j(\bar{E}, \bar{A})$ *for all* j, $1 \leqq j \leqq n$.

Proof. First, suppose that $(\bar{E}, \bar{A}, \bar{B}) \in \overline{O_\eta(E, A, B)}$. The existence of matrices X, Y satisfying condition (i) follows from Lemma 4.1, and the uniqueness was established in Remark 4.2. Clearly, $(\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)}$, and so condition (ii) follows from Theorem 3.7.

Conversely, suppose that conditions (i) and (ii) are satisfied. By Theorem 3.7, condition (ii) implies that $(\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)}$. Then it follows from Lemma 3.2 that $\bar{A}_\alpha \in \overline{O_\gamma(A_\alpha)}$ for any $\alpha \in K$ such that $\det(\alpha E - A) \neq 0$. (Recall that γ denotes the conjugation action on $K^{n\times n}$.) Moreover, it follows from Remark 4.2 that $XA_\alpha = \bar{A}_\alpha X$ and $XB_\alpha = \bar{B}_\alpha$. By Lemma 4.4 of [13], there exists a sequence (R_k) in $\mathrm{Gl}_n(K)$ such that

$$\lim_{k\to\infty} R_k A_\alpha R_k^{-1} = \bar{A}_\alpha \quad \text{and} \quad \lim_{k\to\infty} R_k = X. \tag{4.14}$$

Define $L_k \in \mathrm{Gl}_n(K)$ by

$$L_k = (\alpha\bar{E} - \bar{A})R_k(\alpha E - A)^{-1}. \tag{4.15}$$

It follows from (4.14) that

$$\lim_{k\to\infty} L_k E R_k^{-1} = \bar{E}. \tag{4.16}$$

On the other hand, the definition (4.15) of L_k implies that

$$L_k(\alpha E - A)R_k^{-1} = \alpha\bar{E} - \bar{A}. \tag{4.17}$$

This, together with (4.16), implies that

$$\lim_{k\to\infty} L_k A R_k^{-1} = \bar{A}. \tag{4.18}$$

Because $\lim_{k\to\infty} R_k = X$ and, by (4.12), $XB_\alpha = \bar{B}_\alpha$, we have

$$\lim_{k\to\infty} R_k(\alpha E - A)^{-1}B = (\alpha\bar{E} - \bar{A})^{-1}\bar{B}.$$

It follows from (4.15) that

$$\lim_{k\to\infty} L_k B = \lim_{k\to\infty} (\alpha\bar{E} - \bar{A}) R_k (\alpha E - A)^{-1} B = \bar{B}.$$

Therefore $(\bar{E}, \bar{A}, \bar{B}) \in \overline{O_\eta(E, A, B)}$. □

By duality, we obtain an analogous result for systems of the form

$$E\dot{x} = Ax,$$
$$y = Cx \quad \text{where } \det(sE - A) \not\equiv 0.$$

By [19], we see that such a system is observable if and only if the system

$$E^t\dot{x} = A^t x + C^t u$$

is controllable. Applying Theorem 4.3 to the system (E^t, A^t, C^t), we obtain Corollary 4.4.

COROLLARY 4.4. *Suppose that $(\bar{E}, \bar{A})$ and (E, A) are regular pencils and suppose that (E, A, C) is observable. Let ψ be the action of $\mathrm{Gl}_n(K) \times \mathrm{Gl}_n(K)$ on matrix triples given by*

$$(L, R) \cdot (E, A, C) = (LER^{-1}, LAR^{-1}, CR^{-1}).$$

Then $(\bar{E}, \bar{A}, \bar{C}) \in \overline{O_\psi(E, A, C)}$ if and only if the following two conditions are satisfied:

(i) *There exists a unique pair of matrices U, $W \in K^{n\times n}$ such that $EW = U\bar{E}$, $AW = U\bar{A}$, and $\bar{C} = CW$.*

(ii) *$D_j(E, A)$ divides $D_j(\bar{E}, \bar{A})$ for all j, $1 \leqq j \leqq n$.*

In the regular case where $E = \bar{E} = I_n$, condition (i) coincides with the conclusion of Theorem 4.1 in [13]. Thus Theorem 4.3 yields a straight generalization of the main result of Khadr and Martin. The following corollary characterizes those systems $(E, A, B) \in S(n, m)$ that have a closed η-orbit, i.e., those that degenerate only to equivalent systems.

COROLLARY 4.5. *For any $(E, A, B) \in S(n, m)$, the orbit $O_\eta(E, A, B)$ is closed in $S(n, m)$ if and only if (E, A, B) is equivalent to a system $(\hat{E}, \hat{A}, 0)$ with*

$$\hat{E} = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \hat{A} = \begin{bmatrix} S & 0 \\ 0 & I_{n-r} \end{bmatrix}$$

where S is semisimple.

Proof. If $O_\eta(E, A, B)$ is closed, then B is necessarily zero since every system degenerates to one with $B = 0$ via

$$\lim_{\varepsilon\to 0} \varepsilon I \cdot (E, A, B) \cdot (\varepsilon I)^{-1} = (E, A, 0).$$

A system $(E, A, 0) \in S(n, m)$ has a closed η-orbit if and only if $O_\varepsilon(E, A)$ is closed. Thus the corollary follows from Remark 3.9. □

In the following proposition we prove that every noncontrollable system in $S(n, m)$ is the degeneration of a controllable system, i.e.,

$$S(n, m) = \bigcup_{(E,A,B)\in S^c(n,m)} \overline{O_\eta(E, A, B)},$$

whereas a controllable system cannot degenerate to a controllable system.

PROPOSITION 4.6.

$$S(n, m) \setminus S^c(n, m) = \bigcup_{(E,A,B)\in S^c(n,m)} \partial O_\eta(E, A, B). \tag{4.19}$$

Proof. Since η is an algebraic $\mathrm{Gl}_n(K)$-action, orbits in the boundary of $O_\eta(E, A, B)$ have dimension strictly less than dim $O_\eta(E, A, B)$; see [11, Lemma 8.3]. But we have seen in Proposition 2.2 that orbits of controllable systems all have the same dimension. Therefore

$$S(n, m)\backslash S^c(n, m) \supset \bigcup_{(E,A,B)\in S^c(n,m)} \partial O_\eta(E, A, B).$$

To prove the opposite inclusion, let $(\bar{E}, \bar{A}, \bar{B})$ be an arbitrary element of $S(n, m)\backslash S^c(n, m)$ and choose $\alpha \in K$ such that $\det(\alpha\bar{E} - \bar{A}) \neq 0$. By Theorem 5.1 of [13] there exists a state-space system $(\hat{A}, \hat{B}) \in L^c(n, m)$ such that

$$\rho_\alpha(\bar{E}, \bar{A}, \bar{B}) = (\bar{A}_\alpha, \bar{B}_\alpha) \in \overline{O_\sigma(\hat{A}, \hat{B})}.$$

By Proposition 1.2, there exists a system (E, A, B) in $S^c_\alpha(n, m)$ with $\rho_\alpha(E, A, B) = (A_\alpha, B_\alpha) = (\hat{A}, \hat{B})$, and by Lemma 3.1 the η-orbit of the system (E, A, B) contains $(\bar{E}, \bar{A}, \bar{B})$ in its closure. □

Note that this result is stronger than the statement that $S^c(n, m)$ is dense in $S(n, m)$. The closures of orbits of an η-invariant open and dense subset of $S(n, m)$ do not necessarily cover $S(n, m)$. For instance, the open and dense subset $S_{\mathrm{reg}}(n, m)$ of all regular systems is closed with respect to orbit closure, since no regular system can degenerate to a singular system in $S(n, m)$ (a consequence of (3.16)). For a given noncontrollable system $(\bar{E}, \bar{A}, \bar{B})$ in $S(n, m)$, there will, in general, exist more than one orbit $O_\eta(E, A, B) \subset S^c(n, m)$ that contains $(\bar{E}, \bar{A}, \bar{B})$ in its closure. However, in the single input case $(m = 1)$, this orbit is uniquely determined. A proof of this fact is indicated in the following example.

Example 4.7. Let $m = 1$ and suppose (without loss of generality) that the pencil $(\bar{E}, \bar{A})$ is in standard form:

$$\bar{E} = \begin{bmatrix} I_r & 0 \\ 0 & \bar{N} \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} \bar{J} & 0 \\ 0 & I_{n-r} \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} \bar{b}_1 \\ \bar{b}_2 \end{bmatrix} \begin{matrix} \}r \\ \}n-r \end{matrix}$$

where $r = \deg \det(s\bar{E} - \bar{A})$ and $\bar{N}$, $\bar{J}$ are in Jordan normal form and $\bar{N}$ is nilpotent. Let

$$\bar{J} = \bigoplus_{i=1}^{s} J(\lambda_i, p_i)$$

where $J(\lambda_i, p_i)$ is an $r_i \times r_i$ Jordan matrix with eigenvalue λ_i and block structure p_i; (see (3.9)). Let $O_\eta(E, A, B)$ be an orbit in $S^c(n, 1)$ that contains $(\bar{E}, \bar{A}, \bar{B})$ in its closure. Again we may assume that (E, A) is in standard form with N and J in Jordan normal form. Since (N, b_2) and (J, b_1) are both controllable (1.9), it follows that N has only one nilpotent Jordan block $J(0, (n-r))$ and J has only one Jordan block $J(\lambda_i, (r_i))$ for each eigenvalue λ_i (for ease of notation, we let (a) denote the partition $(a, 0, \cdots, 0)$). Thus the orbit $O_\varepsilon(E, A)$ is uniquely determined by $(\bar{E}, \bar{A})$.

Using [4, VIII.2], it is not difficult to show that any two controllable systems (E, A, B) and (E, A, B') are η-equivalent. Thus we may choose B to be the column matrix with every entry equal to one. Thus, for any $(\bar{E}, \bar{A}, \bar{B}) \in S(n, 1)$, there exists a unique (and easily constructible) orbit $O_\eta(E, A, B)$ in $S^c(n, 1)$ with $(\bar{E}, \bar{A}, \bar{B}) \in \overline{O_\eta(E, A, B)}$.

In the multi-input case, the determination of all η-orbits whose closure contains a given noncontrollable system $(\bar{E}, \bar{A}, \bar{B})$ is an open problem. It is not even known under which conditions the number of such orbits is finite.

In the following examples, we consider the reverse situation: for a given controllable system (E, A, B), we seek to determine all the η-orbits in the closure of $O_\eta(E, A, B)$.

Example 4.8. Consider $\overline{O_\eta(E, A, B)}$, where

$$E = \begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} J_2 & 0 \\ 0 & I_2 \end{bmatrix}, \quad B = \begin{bmatrix} I_2 \\ I_2 \end{bmatrix}.$$

If $(\bar{E}, \bar{A}, \bar{B}) \in \overline{O_\eta(E, A, B)}$, then $(\bar{E}, \bar{A}) \in \overline{O_\varepsilon(E, A)}$. Thus, from Fig. 1, we see that, up to equivalence, there are two choices for $(\bar{E}, \bar{A})$:

(a) $(\bar{E}, \bar{A}) = (E, A)$.

(b) $(\bar{E}, \bar{A})$ is the pair at the bottom of Fig. 1.

Since (E, A, B) is controllable, the possibilities for $\bar{B}$ will be determined by the matrices $X, Y \in K^{n\times n}$ satisfying $YE = \bar{E}X$ and $YA = \bar{A}X$. In case (a), a straightforward computation shows that X and Y must be of the form:

$$X = Y = \left[\begin{array}{cc|c} a & b & 0 \\ 0 & a & \\ \hline & 0 & s \quad t \\ & & u \quad v \end{array}\right], \qquad a, b, s, t, u, v \in K. \tag{4.20}$$

Hence the systems $(E, A, \bar{B})$ in $\overline{O_\eta(E, A, B)}$ are those for which $\bar{b}_{21} = 0$ and $\bar{b}_{11} = \bar{b}_{22}$. Since left multiplication of $\bar{B}$ by nonsingular matrices Y of the form (4.20) yields a system $(E, A, Y\bar{B})$ in the same η-orbit, we find that exactly 12 η-orbits of the form $O_\eta(E, A, \bar{B})$ lie in $\overline{O_\eta(E, A, B)}$, namely, those determined by

$$\bar{B} = \begin{bmatrix} \bar{B}_1 \\ \bar{B}_2 \end{bmatrix}$$

with

$$\bar{B}_1 = I_2, J_2, \text{or } 0 \quad \text{and} \quad \bar{B}_2 = I_2, J_2, 0, \text{or} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.21}$$

Now consider case (b), where $\bar{B} = YB$ and

$$YE = \bar{E}X \quad \text{and} \quad YA = \bar{A}X. \tag{4.22}$$

Then X, Y must be of the form

$$X = Y = \left[\begin{array}{cc|c} 0 & a & 0 \\ 0 & b & \\ \hline & 0 & s \quad t \\ & & u \quad v \end{array}\right], \qquad a, b, s, t, u, v \in K. \tag{4.23}$$

It follows that $\bar{B}$ may be any matrix in $K^{4\times 2}$ satisfying $b_{11} = b_{21} = 0$. Two such systems $(\bar{E}, \bar{A}, \bar{B})$ and $(\bar{E}, \bar{A}, \bar{B}')$ are equivalent if and only if there are nonsingular matrices L, R such that

$$L\bar{E} = \bar{E}R, \quad L\bar{A} = \bar{A}R, \quad \bar{B}' = L\bar{B}. \tag{4.24}$$

It follows that

$$L = R = \begin{bmatrix} S & 0 \\ 0 & T \end{bmatrix} \quad \text{for some } S, T \in \mathrm{Gl}_2(K). \tag{4.25}$$

Then there are eight orbits of the form $O_\eta(\bar{E}, \bar{A}, \bar{B})$ in $\overline{O_\eta(E, A, B)}$: those for which

$$\bar{B} = \begin{bmatrix} \bar{B}_1 \\ \bar{B}_2 \end{bmatrix},$$

where

$$\bar{B}_1 = J_2 \text{ or } 0 \quad \text{and} \quad \bar{B}_2 = I_2, J_2, 0, \quad \text{or} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.26}$$

Although there are only 20 orbits in the closure of the η-orbits in Example 4.8, we see in the next example that there can be infinitely many orbits in the closure of an η-orbit, at least in the case $m \geqq 2$.

Example 4.9. Suppose $n \geqq 1$ and $m \geqq 2$. Let $E = I_n$, let $A = \operatorname{diag}(\alpha_1, \cdots, \alpha_n)$ such that $\alpha_1 = \alpha_2 \in K$ and $\alpha_i \neq \alpha_j$ for all $i \neq j$, $i, j \geqq 2$, and let

$$B = \begin{bmatrix} B_1 \\ \vdots \\ B_n \end{bmatrix}$$

where $B \in K^{n \times m}$ is any matrix such that (I, A, B) is controllable (i.e., all rows B_i of B are nonzero and B_1 and B_2 are independent). By Theorem 4.3, we know that $(I, A, \bar{B}) \in \overline{O_\eta(I, A, B)}$ if there is a matrix $X \in K^{n \times n}$ such that $XA = AX$ and $\bar{B} = XB$. Consider the family of $n \times n$ matrices

$$X(\varepsilon) = \left[\begin{array}{cc|c} 1 & \varepsilon & 0 \\ 0 & 0 & \\ \hline & 0 & 0 \end{array}\right], \qquad \varepsilon \in K,$$

and the resulting family of systems $(I, A, B(\varepsilon))$, $B(\varepsilon) = X(\varepsilon)B$, in $\overline{O_\eta(I, A, B)}$. Two of these systems, $(I, A, B(\varepsilon))$ and $(I, A, B(\varepsilon'))$, are η-equivalent if and only if there exists $T \in \mathrm{Gl}_n(k)$ such that $TA = AT$ and $B(\varepsilon') = TB(\varepsilon)$. If $TA = AT$, then T is of the form

$$T = T_2 \oplus D \text{ where } T_2 \in \mathrm{Gl}_2(K) \text{ and } D \in \mathrm{Gl}_{n-2}(K) \text{ is diagonal.}$$

Hence

$$TB(\varepsilon) = \begin{bmatrix} t_{11}(B_1 + \varepsilon B_2) \\ t_{21}(B_1 + \varepsilon B_2) \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

If $TB(\varepsilon) = B(\varepsilon')$, then $t_{21} = 0$ and

$$t_{11}(B_1 + \varepsilon B_2) = B_1 + \varepsilon' B_2.$$

Since B_1 and B_2 are independent, it follows that $t_{11} = 1$ and $\varepsilon = \varepsilon'$. Therefore, the orbits $\{O_\eta(I, A, B(\varepsilon))\}_{\varepsilon} \in K$ are distinct and so $O_\eta(I, A, B)$ contains an infinite number of orbits.

To date, there is no general description of the η-orbit boundaries that contain only finitely many orbits.

REFERENCES

[1] D. COBB, *Descriptor variables and generalized singularly perturbed systems: a geometric approach*, Ph.D. dissertation, Department of Electrical Engineering, University of Illinois, Urbana, IL, 1980.

[2] ———, *Fundamental properties of the manifold of singular and regular linear systems*, J. Math. Anal. Appl., 120 (1986), pp. 328-353.

[3] ———, *Controllability, observability, and duality in singular systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 1076-1082.

[4] F. R. GANTMACHER, *The Theory of Matrices*, Vol. I and II, Chelsea, New York, 1960.

[5] M. GERSTENHABER, *On nilalgebras and linear varieties of nilpotent matrices*, III, Ann. of Math., 70 (1959), pp. 167-205.

[6] H. GLÜSING-LÜERSSEN AND D. HINRICHSEN, *A Jordan control canonical form for singular systems*, Internat. J. Control, 48 (1988), pp. 1769-1785.
[7] M. HAZEWINKEL AND R. KALMAN, *On Invariants, Canonical Forms and Moduli for Linear Constant, Finite-Dimensional, Dynamical Systems*, Lecture Notes in Economics—Math. System Theory 131, Springer-Verlag, New York, 1976, pp. 48-60.
[8] U. HELMKE AND D. HINRICHSEN, *Canonical forms and orbit spaces of linear systems*, IMA J. Math. Control Inform., 3 (1986), pp. 167-184.
[9] U. HELMKE AND M. A. SHAYMAN, *Topology of the orbit space of generalized linear systems*, preprint.
[10] D. HINRICHSEN AND A. J. PRITCHARD, *Stability radii of linear systems*, Systems Control Lett., 7 (1986), pp. 1-10.
[11] J. E. HUMPHREYS, *Linear Algebraic Groups*, Springer-Verlag, Berlin, Heidelberg, New York, 1975.
[12] I. M. JAMES, *On category, in the sense of Liusternik-Schnirelmann*, Topology, 17 (1978), pp. 341-348.
[13] A. S. KHADR AND C. F. MARTIN, *On the* Gl(n) *action on linear systems: the orbit closure problem*, in Algebraic and Geometric Methods in Linear System Theory, Byrnes and Martin, eds., American Mathematical Society, Providence, RI, 1980.
[14] D. MUMFORD AND J. FOGARTY, *Geometric Invariant Theory*, Ergeb. Math. 34, Springer-Verlag, Berlin, New York, 1982.
[15] A. C. PUGH, G. E. HAYTON, AND P. FRETWELL, *Transformations of matrix pencils and implications in linear systems theory*, Internat. J. Control, 45 (1987), pp. 529-548.
[16] H. H. ROSENBROCK, *Structural properties of linear dynamical systems*, Internat. J. Control, 20 (1974), pp. 191-202.
[17] A. TANNENBAUM, *Invariance and System Theory: Algebraic and Geometric Aspects*, Lecture Notes in Mathematics, 845, Springer-Verlag, Berlin, New York, 1981.
[18] G. VERGHESE, B. C. LEVY, AND T. KAILATH, *A generalized state-space for singular systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 811-831.
[19] E. L. YIP AND R. F. SINCOVEC, *Solvability, controllability, and observability of continuous descriptor systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 702-707.
[20] K. D. YOUNG, P. V. KOKOTOVIC, AND V. I. UTKIN, *A singular perturbation analysis of high-gain feedback systems*, IEEE Trans. Automat. Control, 22 (1977), pp. 931-937.
[21] Z. ZHOU, M. A. SHAYMAN, AND T. J. TARN, *Singular systems: a new approach in the time domain*, IEEE Trans. Automat. Control, 32 (1987), pp. 42-50.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 624-652, May 1990

009

ON DIFFERENTIAL GAMES OF FIXED DURATION WITH PHASE COORDINATE RESTRICTIONS ON ONE PLAYER*

K. HAJI-GHASSEMI†

Abstract. This paper considers differential games of fixed duration in which state constraints described by a given closed set **E** are imposed on one of the players. Using Berkovitz's definition of a game, the existence of the value is obtained first and, under mild conditions, the existence of saddle points. Sufficient conditions are then given for the value to be continuous or Lipschitz continuous.

Key words. differential games, phase restrictions, value, saddle point

AMS(MOS) subject classification. 90D25

Introduction. Consider a differential game of fixed duration with terminal time T, and dynamics:

$$\begin{aligned} &\dot{x} = f(t, x, y, z), \\ &x \in \mathbb{R}^n, \qquad y \in Y, \qquad z \in Z, \end{aligned} \tag{0.1}$$

where Y and Z are compact subsects of $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Using the definition of a game according to Berkovitz, [1], the set-valued payoff, for a game with initial point (t_0, x_0), is defined by

$$\mathbf{P}[t_0, x_0, \Gamma, \Delta] = \{g(\varphi[\mathbf{T}]) \colon \varphi \in \Phi[\cdot, t_0, x_0, \Gamma, \Delta]\},$$

where Γ and Δ are strategies of the first and the second player, respectively, and $\Phi[\cdot, t_0, x_0, \Gamma, \Delta]$ is the set of all motions resulting from (Γ, Δ) (see [1] for definitions), and $g : \mathbb{R}^n \to \mathbb{R}$ is a given function.

Let **E** be a given closed set in $\mathbb{R}^n$. Suppose that in the above game we impose the additional restriction that one of the players, say the first player (maximizer), must choose his strategies in such a way that all resulting motions lie in **E**. Such restrictions can be imposed by defining the payoff through, in place of g, an extended real valued function $V(\cdot)$ defined by

$$V(\varphi) = \begin{cases} g(\varphi[T]) & \text{if } \varphi[t] \in \mathbf{E}\ \forall t \in [t_0, T] \\ -\infty & \text{otherwise.} \end{cases} \tag{0.2}$$

Note. If player II (minimizer) were chosen as the restricted player, one would replace "$-\infty$" by "$+\infty$" in the definition of $V(\varphi)$.

Games with phase coordinate restrictions as above have been considered previously by Friedman ([4], [5]), Scalzo ([8]), Subbotin [9], and Zaremba ([10], [11]). Each author, using a particular definition of a game, establishes the existence ([4], [5], [8], [9], [10], [11]) and continuity (or Lipschitz continuity) ([4], [5], [8]) of the value. In [4] and [5] it is also shown, under assumptions which guarantee the existence and the continuity of the value, that saddle points exist if one restricts the class of strategies appropriately. These are termed "X-saddle points." In this paper, using Berkovitz's definition of a game, introduced in [1], we obtain sharper results on the existence of value and saddle points and the regularity of the value for the games described above.

* Received by the editors October 17, 1988; accepted for publication (in revised form) May 26, 1989.

† Department of Mathematics and Statistics, Miami University, Oxford Ohio 45056. The research for this work was performed while the author was a graduate student at Purdue University. During the preparation of this work, the author was supported by a David Ross grant from Purdue University and, in part, by National Science Foundation grant DMS-8700813.

It is shown, in § 1, that the techniques of [1], slightly extended, give the existence of value in games where the restricted phase set is any closed subset of $\mathbb{R}^n$. In § 2, we show that, in general, saddle points exist in such games if the strategies of player I are limited to a smaller class. But, if the value is assumed to be always finite (but not necessarily continuous), then saddle points do exist without restricting the class of strategies. In § 3, which comprises a large part of the paper, first the regularity properties of the value are examined for an especially simple type of phase set, namely, a half-space: $\{x \in \mathbb{R}^n: x^n \geqq 0\}$. This case occurs frequently in examples. Conditions are given ensuring that the value is always greater than $-\infty$, continuous, or Lipschitz continuous. Here Theorem 3.1 is of the same nature as the existing results of Friedman and Scalzo ([4], [5], [8]), but with much less restrictive hypotheses (cf., for example, [5, Thms. 8.1 and 8.3]). Then extensions to more general sets are briefly commented on. We conclude by showing that if $\mathbf{E}$ is any closed set, continuity of the value of $\partial\mathbf{E}$ implies continuity on $\mathbf{E}$.

Throughout this work, unless otherwise stated, we assume the following concerning the data of the problems.

H1. (i) The function f in (0.1) is continuous on $[0, \mathbf{T}] \times \mathbb{R}^n \times Y \times Z$.

(ii) There is a constant $K > 0$ such that

$$|f(t, x, y, z) - f(t, \bar{x}, y, z)| \leqq K|x - \bar{x}|$$

for all $t \in [0, T]$, $x, \bar{x} \in \mathbb{R}^n$, $y \in Y$, $z \in Z$.

(iii) The function g, used to define the payoff, is continuous.

(iv) (Isaacs' condition). For any $(t, x) \in [0, \mathbf{T}] \times \mathbf{E}$, and $s \in \mathbb{R}^n$

$$\max_y \min_z \langle s, f(t, x, y, z)\rangle = \min_z \max_y \langle s, f(t, x, y, z)\rangle.$$

(v) The phase set $\mathbf{E}$ is closed.

Remarks. 1. For a discussion of the Isaacs' condition, see [1, § 9]. 2. In (ii), it suffices, for our results, that f be locally Lipschitz in x.

We use the concepts and notations introduced in [1] throughout the paper. We assume, therefore, that the reader is familiar with them. We will denote the Euclidean norm of $x \in \mathbb{R}^n$ by $|x|$, $\{x \in \mathbb{R}^n: |x - a| \leqq R\}$ by $B_R(a)$, and if $\mathbf{F} \subset \mathbb{R}^n$, $\inf\{|x - f|: f \in \mathbf{F}\}$ by $d(x, \mathbf{F})$.

1. Existence of the value. The main result of this section is Theorem 1.1. It asserts the existence of the value in games described by (0.1) and (0.2). We lead to it through several lemmas. We assume **H1** throughout this section.

LEMMA 1.1. *Let Δ be a strategy of the second player defined on $[t_0, \mathbf{T}]$, $t_0 \in [0, \mathbf{T}]$, such that for some $c \in \mathbb{R}$, and X, a compact subset of $\mathbf{E}$,*

$$\varphi \in \Phi[\cdot, t_0, x_0, \Delta], \quad x_0 \in X \Rightarrow V(\varphi) < c. \tag{1.1}$$

Then there exists an $\varepsilon > 0$ such that for any $x_0 \in X$ and $\Phi \in \Phi[\cdot, t_0, x_0, \Delta]$,

$$g(\varphi[\mathbf{T}]) \geqq c \Rightarrow \exists \hat{t} \in [t_0, \mathbf{T}] \quad \text{such that } d(\varphi[\hat{t}], \mathbf{E}) \geqq \varepsilon. \tag{1.2}$$

Proof. If the conclusion were false, then there would exist sequences $\{x_n\}$ in X, $\varepsilon_n \to 0$, and $\varphi_n \in \Phi[\cdot, t_0, x_n, \Delta]$ such that

$$g(\varphi_n[\mathbf{T}]) \geqq c \quad \text{and} \quad d(\varphi_n[t], \mathbf{E}) < \varepsilon_n, \quad \forall t \in [t_0, \mathbf{T}].$$

Since by [1, Lemma 6.2] $\cup\{\Phi[\cdot, t_0, x_n, \Delta]: n = 1, 2, \cdots\}$ is compact in the space of continuous functions on $[t_0, \mathbf{T}]$ with the uniform topology, we may assume, by taking

a subsequence, that there exist an $x_0 \in X$, and a motion $\varphi \in \Phi[\cdot, t_0, x_0, \Delta]$ such that as $n \to \infty$, $x_n \to x_0$, and $\varphi_n \to \varphi$, uniformly on $[t_0, T]$. Hence, for any $t \in [t_0, T]$,

$$d(\varphi[t], \mathbf{E}) = \lim_{n\to\infty} d(\varphi_n[t], \mathbf{E}) = 0.$$

Since $\mathbf{E}$ is closed, this gives $\varphi[t] \in \mathbf{E}$ for all $t \in [t_0, T]$. Therefore, by the definition of V and the continuity of g,

$$V(\varphi) = g(\varphi[T]) = \lim_{n\to\infty} g(\varphi_n[T]) \geqq c.$$

This contradicts (1.1). □

LEMMA 1.2. *Let $\alpha \in \mathbb{R}$. The set $D(\alpha) = \{(t, x): w^+(t, x) < \alpha\}$ is open. (In particular, $w^+(\cdot, \cdot)$ is upper semicontinuous.)*

Proof. Let $(t_0, x_0) \in D(\alpha)$ and $0 < \delta < (\alpha - w^+(t_0, x_0))/2$. Take $c = \alpha - \delta$. Then $w^+(t_0, x_0) < c$. Hence, by the definition of $w^+(t_0, x_0)$, there exists a strategy Δ such that for any $\varphi \in \Phi[\cdot, t_0, x_0, \Delta]$, we have $V(\varphi) < c$. Let $\varepsilon > 0$ be as in Lemma 1.1. Let $t_1 \in [0, \mathbf{T}]$, and define $s: [t_1, \mathbf{T}] \to [t_0, \mathbf{T}]$ by $s(t) = t_0 + (\mathbf{T} - t_0)(t - t_1)/(\mathbf{T} - t_1)$. Let $x_1 \in E$. Let $\Theta = \Theta(t_0, x_0, t_1, x_1)$ be the one-one onto map from the set of all strategies in the game with initial point (t_1, x_1) to those with initial point (t_0, x_0) defined using $s(\cdot)$ as in [1, p. 181]. By [1, Lemma 6.5], for every motion $\varphi \in \Phi[\cdot, t_1, x_1, \Theta^{-1}\Delta]$, there exists a motion $\bar{\varphi} \in \Phi[\cdot, t_0, x_0, \Delta]$ such that

$$\max \{|\varphi[t] - \bar{\varphi}[s(t)]|: t \in [t_1, T]\} \leqq \eta(\rho) \tag{1.3}$$

where $\rho \equiv |t_0 - t_1| + |x_0 - x_1|$, and $\eta(\rho) \to 0$ as $\rho \to 0$.

Let G be a bounded neighborhood of (t_0, x_0). It follows from **H1** that there exists an $R > 0$ such that any motion φ with its initial point in G satisfies $|\varphi[t]| \leqq R$. By the continuity of $g(\cdot)$, there exists $\sigma_1 > 0$ such that

$$0 < \sigma_1 < \varepsilon/2, \quad \text{and } |g(x) - g(x')| < \delta/2, \quad \forall x, x' \in B_R(0), |x - x'| < \sigma_1. \tag{1.4}$$

Let $\sigma_2 > 0$ be such that $\rho < \sigma_2 \Rightarrow \eta(\rho) < \sigma_1$. Now suppose $\rho < \sigma_2$, and let $\varphi \in \Phi[\cdot, t_1, x_1, \Theta^{-1}\Delta]$. Let $\bar{\varphi} \in \Phi[\cdot, t_0, x_0, \Delta]$ such that (1.3) holds. Then either there exists a $\bar{t} \in [t_1, \mathbf{T}]$ with $\varphi[\bar{t}] \notin \mathbf{E}$, in which case, $V(\varphi) = -\infty$, or $\varphi[t] \in \mathbf{E}$ for all $t \in [t_1, \mathbf{T}]$, in which case, by (1.3) and (1.4), $d(\bar{\varphi}[t], \mathbf{E}) < \varepsilon/2$, for all $t \in [t_1, \mathbf{T}]$. Hence, by the choice of ε, $g(\bar{\varphi}[T]) < c$. Now, by (1.4) and the choice of δ, $g(\varphi[T]) < c + \delta/2 < \alpha$. Thus in either case, $V(\varphi) < \alpha - \delta/2$. Since $\varphi \in \Phi[\cdot, t_1, x_1, \Theta^{-1}\Delta]$ was chosen arbitrarily, we have

$$w^+(t_1, x_1) \leqq \sup_{\Gamma} P[t_1, x_1, \Gamma, \Delta] < \alpha.$$

We have shown that $|t_0 - t_1| + |x_0 - x_1| < \sigma_2 \Rightarrow (t_1, x_1) \in D(\alpha)$. Therefore $D(\alpha)$ is open. □

COROLLARY. *Let $(t_0, x_0) \in [0, T] \times E$, $v^0 = w^+(t_0, x_0)$, and*

$$C^+(v^0) = \{(t, x): w^+(t, x) \geqq v^0\}.$$

Then $C^+(v^0) \neq \varnothing$, and is closed. Moreover if $v^0 \neq -\infty$, then any $(t, x) \in C^+(v^0)$ has $x \in \mathbf{E}$.

Proof. $C^+(v^0) \neq \varnothing$ since $(t_0, x_0) \in C^+(v^0)$. The last statement is clear. Closure follows from Lemma 1.2. □

Lemma 1.4, below, is the analogue of [1, Lemma 8.3]; i.e., it states that the sets $C^+(\alpha)$, above are "u-stable"; see Remark 1 below. For future reference, we separate out the main argument of its proof in the next lemma. The proof is essentially the same as that given in [3]. We will therefore omit some of the details.

LEMMA 1.3. *Let* $t_1 \in [0, T]$, *and* $X \subset \mathbf{E}$ *be a compact set such that for some real number* c,

$$w^+(t_1, x) < c, \quad \forall x \in X.$$

Then there exists a strategy Δ^*, *defined on* $[t_1, \mathbf{T}]$, *such that*

$$V(\varphi) < c \quad \textit{for all } \varphi \in \cup\{\Phi[\cdot, t_1, x, \Delta^*]: x \in X\}.$$

Proof. Since X is compact and, by Lemma 1.2, w^+ is upper semicontinuous, there exists an $x^* \in X$ such that $w^+(t_1, x^*) = \max\{w^+(t_1, x): x \in X\}$. By assumption $c - w^+(t_1, x^*) > 0$. Let $\sigma \in (0, c - w^-(t_1, x^*))$. Then $w^+(t_1, x) < c - \sigma$ for all $x \in X$. Therefore, by the definition of w^+, for every $x \in X$, there exists a strategy $\Delta(x)$ such that

$$\varphi \in \Phi[\cdot, t_1, x, \Delta(x)] \Rightarrow V(\varphi) < c - \sigma. \tag{1.5}$$

Now, by Lemma 1.1, there exists an $\varepsilon > 0$ such that for any $\varphi \in \Phi[\cdot, t_1, x, \Delta(x)]$,

$$g(\varphi[\mathbf{T}]) \geqq c - \sigma \Rightarrow d(\varphi[\bar{t}], \mathbf{E}) \geqq \varepsilon. \tag{1.6}$$

It follows from [1, Lemma 6.5] and the continuity of g that there exists a $\delta(x)$ such that if $\bar{x} \in X$, $|x - \bar{x}| < \delta(x)$ then for all $\bar{\varphi} \in \Phi[\cdot, t_1, \bar{x}, \Theta\Delta(x)]$, there exits a $\varphi \in \Phi[\cdot, t_1, x, \Delta(x)]$ (with $\Theta = \Theta(t_1, \bar{x}, t_1, x)$ as in Lemma 1.2) such that

$$\text{(a)} \quad |\varphi[t] - \bar{\varphi}[t]| < \varepsilon/2, \quad \forall t \in [t_1, \mathbf{T}], \tag{1.7}$$

$$\text{(b)} \quad |g(\varphi[\mathbf{T}]) - g(\bar{\varphi}[\mathbf{T}])| < \sigma/2.$$

By (1.5), (1.6), and (1.7), either $g(\bar{\varphi}[\mathbf{T}]) < c - \sigma/2$ or for some $\bar{t} \in [t_1, T]$, $d(\bar{\varphi}[\bar{t}], \mathbf{E}) \geqq \varepsilon/2$. In either case $V(\bar{\varphi}) < c - \sigma/2$. Thus if $|x - \bar{x}| < \delta(x)$, then

$$V(\bar{\varphi}) < c - \sigma/2, \quad \forall \bar{\varphi} \in \Phi[\cdot, t_1, \bar{x}, \Theta\Delta(x)]. \tag{1.8}$$

Let $B(x)$ be the ball of radius $\delta(x)$ centered at x. Then $X \subset \cup\{B(x): x \in X\}$. Since X is compact, there exist $x_1, \cdots, x_k$ such that $X \subset \cup\{B(x_i): i = 1, \cdots, k\}$. Now we define a strategy Δ^* as follows (see [3] for a more formal definition): if $\Pi_{n,i}$ denotes the nth-stage partition of $\Delta(x_i)$, $i = 1, \cdots, k$, let the nth-stage partition of Δ^* be $\Pi_n = \cup\{\Pi_{n,i}: i = 1, \cdots, k\}$. Given the initial point $x \in X$, let $j = j(x) = \min\{i: 1 \leqq i \leqq k$ and $|x - x_i| < \delta(x_i)\}$. Then Δ^*, at the nth-stage, plays the nth-stage of $\Theta\Delta(x_j)$. Note that this is possible since the partition points of $\Pi_{n,j}$ are among those of Π_n. It follows that if $\varphi \in \cup\{\Phi[\cdot, t_1, x, \Delta^*]: x \in X\}$ then $\varphi \in \Phi[\cdot, t_1, x, \Theta\Delta(x_{j(x)})]$. Therefore, by (1.8), $V(\varphi) < c - \sigma/2 < c$ for all $\varphi \in \cup\{\Phi[\cdot, t_1, x, \Delta^*]: x \in X\}$ as desired. □

LEMMA 1.4. *Let* $\alpha \in \mathbb{R}$, *and* $C^+(\alpha) = \{(t, x): w^+(t, x) \geqq \alpha\}$. *Let* $(\tau, \xi) \in C^+(\alpha)$ *and let* $\tau < t_1 < \mathbf{T}$. *Let* $v(\cdot)$ *be any control of the second player on* $[\tau, \mathbf{T}]$. *Then there exists a relaxed control* η *such that the corresponding relaxed trajectory* $\psi(\cdot)$, *with initial point* (τ, ξ) *satisfies* $(t_1, \psi(t_1)) \in C^+(\alpha)$.

Remark 1. The property described above will be referred to as the "u-stability" (or simply "stability") of the set $C^+(\alpha)$.

Proof. Suppose the assertion were false, then there would exist $t_1 \in (\tau, T)$ and $v(\cdot)$, a control function for player II, defined on $[\tau, \mathbf{T}]$ such that if $\psi(\cdot)$ is a relaxed trajectory corresponding to a relaxed control η, then $(t_1, \psi(t_1)) \notin C^+(\alpha)$. Let Ψ be the set consisting of all points $\psi(t_1)$ for all such ψ. Then Ψ is compact ([1, Lemmas 6.1, 6.2]). Set $X = \Psi \cap \mathbf{E}$. X is compact since $\mathbf{E}$ is closed and Ψ is compact. Also, for all $x \in X$, $(t_1, x) \notin C^+(\alpha)$. Therefore, by the definition of $C^+(\alpha)$, $w^+(t_1, x) < \alpha$ for all $x \in X$. Since w^+ is upper semicontinuous and X is compact, there exists $\alpha' < \alpha$ such that $w^+(t_1, x) < \alpha' < \alpha$, for all $x \in X$. Hence, by Lemma 1.3, there exists a strategy Δ^*, defined on $[\tau, \mathbf{T}]$, such that

$$V(\varphi) < \alpha', \quad \forall \varphi \in \cup\{\Phi[\cdot, t_1, x, \Delta^*]: x \in X\}. \tag{1.9}$$

Let $\hat{\Delta}$ be the concatenation of $v(\cdot)$, on $[\tau, t_1]$, and Δ^*, on $[t_1, \mathbf{T}]$. Then for every $\bar{\varphi} \in \Phi[\cdot, \tau, \xi, \hat{\Delta}]$, there exists a motion $\varphi \in \Phi[\cdot, t_1, \bar{\varphi}[t_1], \Delta^*]$ such that

$$\bar{\varphi}[t] = \varphi[t], \quad \forall t \in [t_1, \mathbf{T}]. \tag{1.10}$$

Hence $V(\bar{\varphi}) \leqq V(\varphi)$. By the definition of $\hat{\Delta}$, and [1, Lemma 6.1], there exists a relaxed trajectory ψ such that $\psi(t) = \bar{\varphi}[t]$ for all $t \in [\tau, t_1]$. Therefore if $\bar{\varphi}[t] \in E$ for all $t \in [\tau, t_1]$, then $\bar{\varphi}[t_1] \in X$. By (1.9), $V(\varphi) < \alpha'$. Using (1.10), we have $V(\bar{\varphi}) < \alpha' < \alpha$. Since $\bar{\varphi} \in \Phi[\cdot, t_1, \xi, \hat{\Delta}]$ was chosen arbitrarily, we conclude

$$w^+(\tau, \xi) \leqq \sup\{V(\bar{\varphi}): \bar{\varphi} \in \Phi[\cdot, \tau, \xi, \hat{\Delta}]\} \leqq \alpha' < \alpha.$$

This contradicts the assumption that $(\tau, \xi) \in C^+(\alpha)$, proving the lemma. □

Remark 2. Lemmas 1.2 and 1.4 have shown that if $C^+(\alpha)$, $\alpha \in \mathbb{R}$ is nonempty, then it is closed and "u-stable." Hence we may define extremal strategies, $\Gamma_e = \Gamma_e(C^+(\alpha))$, as in [1, p. 189].

The next lemma is the analogue of [1, Lemma 10.1]. Note that Lemma 9.1 of [1] is valid in our setting since it is independent of the pay-off. Therefore, using the closure and the u-stability of $C^+(\alpha)$, the same arguments as in [1, Lemma 10.1] prove the next lemma.

LEMMA 1.5. *Let* **H1** *hold and let* $(t_0, x_0) \in [0, \mathbf{T}] \times \mathbf{E}$. *Let* $v^0 = w^+(t_0, x_0)$, *and* $\Gamma_e = \Gamma_e(C^+(v^0))$. *Then for every motion* $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma_e]$,

$$(t, \varphi[t]) \in C^+(v^0), \quad \forall t \in [t_0, \mathbf{T}].$$

It follows from this lemma, therefore, that $V(\varphi) = g(\varphi[\mathbf{T}]) \geqq v^0$, for all $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma_e]$. Now the existence of the value follows easily.

THEOREM 1.1. *Let* **H1** *hold and let* $(t_0, x_0) \in [0, \mathbf{T}] \times \mathbf{E}$. *Then the differential game described by* (0.1), *and* (0.2) *has value* $w(t_0, x_0)$.

Proof. Let $v^0 = w^+(t_0, x_0)$. If $v^0 = -\infty$ then clearly $w^-(t_0, x_0) = -\infty$ also, and the value exists trivially. Suppose $v^0 \neq -\infty$, and let $\Gamma_e = \Gamma_e(C^+(v^0))$. Then by Lemma 1.5, for every motion $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma_e]$, $V(\varphi) \geqq v^0$. Hence,

$$w^-(t_0, x_0) \geqq \inf_{\Delta} P[t_0, x_0, \Gamma_e, \Delta] \geqq v^0 = w^+(t_0, x_0).$$

Since the reverse inequality always holds, we are done. □

Remark 3. It is clear that games with phase restrictions on the minimizer, player II, can be treated similarly. Thus games with phase restrictions on only one of the players, in the sense described above, have value.

We conclude this section with a discussion of the set

$$C^-(\alpha) = \{(t, x): t \in [0, T], x \in E, \text{ and } w^-(t, x) \leqq \alpha\}, \quad \alpha \in \mathbb{R},$$

in games with phase restrictions only on the maximizer. Let $(\tau, \xi) \in C^-(\alpha)$ (some $\alpha \in \mathbb{R}$), and $t_1 \in [\tau, T]$. Let $u(\cdot)$ be a control function of the first player, defined on $[\tau, t_1]$, with the property that for any control function, $v(\cdot)$, of the second player and a solution, φ, of

$$\dot{x}(t) = f(t, x, u(t), v(t)), \quad \text{a.e. } t \in [\tau, t_1], \qquad x(\tau) = \xi, \tag{1.11}$$

φ satisfies $\varphi(t) \in E$ for all $t \in [\tau, t_1]$. Then by arguments similar to those of Lemma 1.4, we can show that there exists a relaxed control ζ such that the corresponding relaxed trajectory ψ satisfies

$$(t_1, \psi(t_1)) \in C^-(\alpha). \tag{1.12}$$

Now if $\xi \in \text{int}(E)$, it follows from **H1** that there is a $\delta = \delta(\xi) > 0$ such that for any control function $v(\cdot)$ and solution φ of (1.11) $\varphi(t) \in E$, $t \in [\tau, \tau + \delta]$. Hence we have the following lemma.

LEMMA 1.6. *Let* $(\tau, \xi) \in C^{-}(\alpha)$, *some* $\alpha \in \mathbb{R}$. *Let* $\xi \in \text{int}(E)$. *Then there exists a* $\delta = \delta(\xi) > 0$ *such that for any* $t_1 \in (\tau, \tau + \delta)$, *and any* $u(\cdot)$, *a control of the first player, defined on* $[\tau, t_1]$, *there exists a relaxed control* ζ *such that* ψ, *the corresponding relaxed trajectory, satisfies* (1.12).

2. Saddle points. In this section, we consider the question of the existence of saddle points in a game of § 1 with initial point (t_0, x_0). If $w(t_0, x_0) = -\infty$, then there exists a Δ^* such that $V(\varphi) = -\infty$ for all $\varphi \in \Phi[\cdot, t_0, x_0, \Delta^*]$. Hence for any Γ, the pair (Γ, Δ^*) forms a saddle point trivially. Hence we will assume, below, that $w(t_0, x_0) > -\infty$. We will first show, in Theorem 2.1, that, in general, saddle points exist if we restrict the class of strategies of the first player, appropriately. We will then prove the main theorem of this section, Theorem 2.2, which states that if $w(t, x) > -\infty$ for all $(t, x) \in [0, \mathbf{T}] \times \mathbf{E}$, (but not necessarily continuous there), then saddle points exist without any restrictions on the class of strategies.

We will use the following notation in this and later sections. For a pair of strategies (Γ, Δ),

1. $\Phi_m(\cdot, t_0, x_0, \Gamma, \Delta)$ will denote the set of all mth-stage trajectories resulting from (Γ, Δ) with initial point (t_0, x_0).
2. $\Phi_m(\cdot, t_0, x_0, \Gamma) = \cup \{\Phi_m(\cdot, t_0, x_0, \Gamma, \Delta): \Delta\}$.
3. $\delta_m(\Gamma) = \sup \{\max_{t \in [t_0, T]} d(\varphi_m(t), E): \varphi_m \in \Phi_m(\cdot, t_0, x_0, \Gamma)\}$.
4. $H_t = \{(t, x): x \in \mathbb{R}^n\}$.

DEFINITION. Let us call a class $\mathbb{\Gamma}$ of strategies Γ of the first player, "*restricted*" if $\delta_m(\Gamma) \to 0$ as $m \to \infty$, *uniformly* in $\Gamma \in \mathbb{\Gamma}$.

THEOREM 2.1. *Let* $(t_0, x_0) \in [0, \mathbf{T}] \times \mathbf{E}$, *with* $w(t_0, x_0) > -\infty$, *and* Γ_0 *be the extremal pointing strategy defined with respect to* $C^{+}(w(t_0, x_0))$. *Then for any restricted class* $\mathbb{\Gamma}$ *of strategies of player I, there exists* Δ_0 *such that*

$$\text{(2.1)} \qquad \text{(a)} \quad P[t_0, x_0, \Gamma_0, \Delta_0] = \{w(t_0, x_0)\},$$

$$\text{(b)} \quad \textit{For all } \Delta \textit{ and all } \Gamma \in \mathbb{\Gamma},$$

$$P[t_0, x_0, \Gamma, \Delta_0] \subseteqq P[t_0, x_0, \Gamma_0, \Delta_0] \subseteqq P[t_0, x_0, \Gamma_0, \Delta].$$

The theorem follows easily from Propositions 2.1 and 2.2 below. We first prove these propositions.

PROPOSITION 2.1. *Let* $\mathbb{\Gamma}^*$ *be the set of all strategies,* Γ, *of player I such that for any* $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma]$, $\varphi[t] \in E$, *for all* $t \in [t_0, \mathbf{T}]$. *Then for any* $\Gamma \in \mathbb{\Gamma}^*$, *we have* $\delta_m(\Gamma) \to 0$ *as* $m \to \infty$.

Proof. Suppose that for some Γ the conclusion were false. Then, by taking a subsequence, we may assume that there exists an $\varepsilon > 0$ such that $\delta_m(\Gamma) > \varepsilon$ for all m. By the definition of $\delta_m(\Gamma)$, there exists a $\varphi_m \in \Phi_m(\cdot, t_0, x_0, \Gamma)$ such that

$$\text{(2.2)} \qquad d(\varphi_m(t_m), \mathbf{E}) > \varepsilon \quad \text{for some } t_m \in [t_0, \mathbf{T}].$$

We may assume, by taking another subsequence if necessary, that there exist a $\bar{t} \in [t_0, T]$ and a $\bar{\varphi} \in \Phi[\cdot, t_0, x_0, \Gamma_0]$ such that

$$\text{(2.3)} \qquad \begin{aligned} &t_m \to \bar{t}, \quad \text{as } m \to \infty, \text{ and} \\ &\varphi_m \to \bar{\varphi}, \quad \text{uniformly on } [t_0, T], \text{ as } m \to \infty. \end{aligned}$$

Assumptions **H1** ensure that $\{\varphi_m\}$ is an equicontinuous sequence. Hence, it follows from (2.2) and (2.3) that $d(\bar{\varphi}[\bar{t}], E) \geqq \varepsilon$, contradicting the hypothesis. □

Remark. In particular, by Lemma 1.5, $\delta_m(\Gamma_0) \to 0$ as $m \to \infty$. It follows that if $\mathbb{\Gamma}$ is a restricted class then $\mathbb{\Gamma} \cup \{\Gamma_0\}$ is also a restricted class. *Hence we will always assume from now on that a restricted class contains* Γ_0. Note that, in a similar way, if $\Gamma \in \mathbb{\Gamma}^*$ and $\mathbb{\Gamma}$ is restricted, then $\mathbb{\Gamma} \cup \{\Gamma\}$ is also restricted.

PROPOSITION 2.2. *Let* $\mathbb{\Gamma}$ *be a restricted class. Then there exists a* Δ_0 *such that*

$$P[t_0, x_0, \Gamma, \Delta_0] \leqq w(t_0, x_0) \text{ for all } \Gamma \in \mathbb{\Gamma},$$

and

$$P[t_0, x_0, \Gamma_0, \Delta_0] = w(t_0, x_0).$$

Proof. The second assertion follows from the first and from Lemma 1.5. We now prove the first assertion. Let $v_0 = w(t_0, x_0) > -\infty$. By definition,

$$v_0 = \inf_{\Delta} \sup_{\Gamma} P[t_0, x_0, \Gamma, \Delta].$$

Hence for every n there exists a Δ_n such that, for every $\Gamma \in \mathbb{\Gamma}$,

$$P[t_0, x_0, \Gamma, \Delta_n] \leqq \sup_{\Gamma} P[t_0, x_0, \Gamma, \Delta_n] < v_0 + 1/n. \tag{2.4}$$

Claim. For every n, there exists an $m(n)$ such that if $m \geqq m(n)$, then for any $\varphi_m \in \bigcup \{\Phi_m(\cdot, t_0, x_0, \Gamma, \Delta_n) : \Gamma \in \mathbb{\Gamma}\}$, we have $g(\varphi_m(T)) < v_0 + 1/n$.

Verification. From (2.4) we have that for all $\varphi \in \Phi[\cdot, t_0, x_0, \Delta_n]$, $V(\varphi) < v_0 + 1/n$. By Lemma 1.1, there exists an $\varepsilon_n > 0$ such that for all $\varphi \in \Phi[\cdot, t_0, x_0, \Delta_n]$,

$$g(\varphi[\mathbf{T}]) \geqq v_0 + 1/n \Rightarrow \exists \bar{t} \in [t_0, \mathbf{T}] \text{ satisfying } d(\varphi[\bar{t}], \mathbf{E}) \geqq \varepsilon_n. \tag{2.5}$$

Since $\mathbb{\Gamma}$ *is restricted, there exists an* $m(n)$ such that

$$m \geqq m(n) \Rightarrow \delta_m(\Gamma) < \varepsilon_n/2, \quad \forall \Gamma \in \mathbb{\Gamma}. \tag{2.6}$$

Hence if $m \geqq m(n)$ and $\varphi_m \in \bigcup \{\Phi_m(\cdot, t_0, x_0, \Gamma, \Delta_n) : \Gamma \in \mathbb{\Gamma}\}$ then,

$$\max \{d(\varphi_m(t), \mathbf{E}) : t \in [t_0, \mathbf{T}]\} \leqq \delta_m(\Gamma) < \varepsilon_n/2. \tag{2.7}$$

Now suppose that the claim is false. Then there exists an n and a sequence $m(i) \to \infty$ as $i \to \infty$, and $\{\varphi_{m(i)}\} \subset \bigcup \{\Phi_{m(i)}(\cdot, t_0, x_0, \Gamma, \Delta_n) : \Gamma \in \mathbb{\Gamma}\}$ such that $g(\varphi_{m(i)}(T)) \geqq v_0 + 1/n$. Since $\{\varphi_{m(i)}\}$ is a sequence of $m(i)$th-stage trajectories of Δ_n, there exists a motion $\varphi \in \Phi[\cdot, t_0, x_0, \Delta_n]$ such that $\varphi_{m(i)} \to \varphi$ uniformly on $[t_0, \mathbf{T}]$, as $i \to \infty$. Hence by the continuity of g, $g(\varphi[\mathbf{T}]) \geqq v_0 + 1/n$. Therefore, by (2.5), there exists a $\bar{t} \in [t_0, T]$ such that $d(\varphi[\bar{t}], \mathbf{E}) \geqq \varepsilon_n$. Since $\varphi_{m(i)}$ converges to φ uniformly, we have

$$d(\varphi_{m(i)}(\bar{t}), \mathbf{E}) \geqq \varepsilon_n/2,$$

for all i sufficiently large. This contradicts (2.7), proving the claim.

Now define Δ_0 as $\{\Delta_{n,m(n)}, \Pi_{n,m(n)}\}$ where $\Pi_{n,i}$ is the ith-stage partition of Δ_n. That is, the nth-stage of Δ_0 is the $m(n)$th-stage of Δ_n. Hence if $\Gamma \in \mathbb{\Gamma}$, $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma, \Delta_0]$ and $\{\varphi_n\}$ is the sequence of nth-stage trajectories converging to φ, then, by the definition of Δ_0, $\varphi_n \in \Phi_{m(n)}(\cdot, t_0, x_0, \Gamma, \Delta_n)$. Therefore, by the above claim, $g(\varphi_n(T)) < v_0 + 1/n$. Hence $V(\varphi) \leqq g(\varphi[T]) \leqq v_0$. Since Γ in $\mathbb{\Gamma}$ and $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma, \Delta_0]$ were chosen arbitrarily, we conclude:

$$P[t_0, x_0, \Gamma, \Delta_0] \leqq v_0 \quad \text{for all } \Gamma \in \mathbb{\Gamma}, \tag{2.8}$$

as desired. □

Proof of Theorem 2.1. As noted above,

$$P[t_0, x_0, \Gamma_0, \Delta_0] = w(t_0, x_0). \tag{2.9}$$

On the other hand, by Lemma 1.5, $w(t_0, x_0) \leqq P[t_0, x_0, \Gamma_0, \Delta]$ for any Δ. Hence, using (2.8) and (2.9), we have

$$P[t_0, x_0, \Gamma, \Delta_0] \leqq P[t_0, x_0, \Gamma_0, \Delta_0] \leqq P[t_0, x_0, \Gamma_0, \Delta]$$

for all strategies Δ and all $\Gamma \in \mathbb{\Gamma}$, as was claimed. □

Before we state and prove Theorem 2.2, let us note that if $\Omega \subset [0, \mathbf{T}] \times \mathbf{E}$ is closed and "u-stable," then Lemmas 9.1 and 10.1 of [1] prove the following lemma.

LEMMA 2.1. *Suppose Ω is closed and u-stable. Let X be a compact subset of $\mathbf{E}$ and let $t \in [0, \mathbf{T}]$ be such that $\Omega \cap H_t \neq \varnothing$. Let Γ_e be the strategy defined extremally with respect to Ω. Then there exists a constant $C > 0$, independent of Ω, such that for any $\varphi \in \cup \{\Phi[\cdot, t, x, \Gamma_e]: x \in X\}$ we have*

$$d((\tau, \varphi[\tau]), \Omega \cap H_\tau) \leqq C d((t, x), \Omega \cap H_t).$$

Proof. If $(t, x) \in \Omega$ then the proof is exactly the same as that of [1, Lemma 10.1]. Otherwise, using the same notation as in [1, Lemma 10.1], we have $k = k(m) = 1$, and

$$\lim_{m \to \infty} \varepsilon_m(\tau_k) = d((t, x), C^+(\alpha) \cap H_t),$$

where $\varepsilon_m(\tau) \equiv d((\tau, \varphi_m(\tau)), C^+(\alpha) \cap H_\tau)$. As in that Lemma, we obtain

$$\varepsilon_m^2(t) \leqq \varepsilon_m^2(\tau_k)\, e^{\beta K} + \mathbf{E}(\delta_m)(e^{\beta K} - 1)/\beta, \quad \forall \tau \in [\tau_k, T],$$

where $\mathbf{E}(\cdot)$ is a function, depending on X and the modulus of continuity of f, such that $\mathbf{E}(r) \to 0$ as $r \to 0$, $K =$ Lipschitz constant of f and β is a constant, which in our setting depends only on f and X. Letting $m \to \infty$, we obtain the desired conclusion with $C = e^{\beta K/2}$. □

We are now ready to state and prove Theorem 2.2.

THEOREM 2.2. *Suppose that $w(t, x) > -\infty$ for all $(t, x) \in [0, \mathbf{T}] \times \mathbf{E}$. Then saddle points exist; i.e., for every $(t_0, x_0) \in [0, T] \times \mathbf{E}$, there exists Δ_0 such that for any Γ and Δ,*

$$P[t_0, x_0, \Gamma, \Delta_0] \leqq P[t_0, x_0, \Gamma_0, \Delta_0] \leqq P[t_0, x_0, \Gamma_0, \Delta],$$

where Γ_0 is the extremal strategy with respect to $C^+(t_0, x_0)$.

Remark. Conditions which guarantee $w(t, x) > -\infty$ for all $(t, x) \in [0, \mathbf{T}] \times \mathbf{E}$ can be given when more information is known about $\mathbf{E}$; see, for example, Proposition 3.1 below.

Proof. It can be easily verified that $w(t, x) > -\infty$ for all $(t, x) \in [0, \mathbf{T}] \times \mathbf{E}$ if and only if $[0, \mathbf{T}] \times \mathbf{E}$ is u-stable. Now, let $(t_0, x_0) \in [0, \mathbf{T}] \times \mathbf{E}$. Let $R > 0$ be such that all trajectories with initial point (t_0, x), x in a bounded neighborhood of x_0, remain in $B_R(0)$ on $[t_0, T]$. Let $M = \max\{|f(t, x, y, z)|: t \in [t_0, T], |x| \leqq R, y \in Y, z \in Z\}$. Let Γ_e denote the extremal strategy defined with respect to $[0, \mathbf{T}] \times \mathbf{E}$. By Lemma 2.1, there exists a constant $C > 0$ such that for all $\varphi \in \cup \{\Phi[\cdot, t_0, x, \Gamma_e]: |x| \leqq R\}$, we have $d(\varphi[t], \mathbf{E}) \leqq C d(x, \mathbf{E})$. Fix $\mathcal{K} > C$. Then it follows from the compactness of $\cup \{\Phi[\cdot, t_0, x, \Gamma_e]: |x| \leqq R\}$ (cf., [1, Lemma 6.1]) that there exists an m^* such that if $m \geqq m^*$ then

$$d(\varphi_m(t), \mathbf{E}) \leqq \mathcal{K} d(x, \mathbf{E}), \quad \text{for all } \varphi_m \in \cup \{\Phi_m(\cdot, t_0, x, \Gamma_e): |x| \leqq R\}. \tag{2.10}$$

Without loss of generality, we will assume that $m^* = 1$.

Now let $\mathbb{\Gamma} = \{\Gamma: \delta_m(\Gamma) \leqq C[\delta_m(\Gamma_0) + M(T - t_0)/m]\}$. Then, by Proposition 2.2, there exists a Δ_0 such that

$$P[t_0, x_0, \Gamma, \Delta_0] \leqq P[t_0, x_0, \Gamma_0, \Delta_0], \quad \text{for all } \Gamma \in \mathbb{\Gamma}. \tag{2.11}$$

Claim. Δ_0 has the desired property; i.e., (2.11) holds for any Γ.

Verification. Suppose that the claim is false. Then there exists a strategy Γ and a motion $\varphi \in \Phi[\cdot\, t_0, x_0, \Gamma, \Delta_0]$ such that $V(\varphi) > v_0$. In particular, $\varphi[t] \in \mathbf{E}$, for all $t \in [t_0, T]$. Let $\{\varphi_m(\cdot)\}$ be the sequence of mth-stage trajectories converging to φ. Let (u_m, v_m) denote the controls of φ_m. We will show that there exists a $\bar{\Gamma} \in \mathbb{\Gamma}$ such that $\varphi \in \Phi[\cdot, t_0, x_0, \bar{\Gamma}, \Delta_0]$. This contradicts (2.11), proving the claim.

Before defining $\bar{\Gamma}$, observe that since $\varphi[t] \in E$, for all $t \in [t_0, T]$, then

$$\max\{d(\varphi_m(t), \mathbf{E}): t \in [t_0, T]\} \to 0 \quad \text{as } m \to \infty.$$

Hence for every m, there exists $n(m) \geqq m$ such that

$$\max\{d(\varphi_n(t), \mathbf{E}): t \in [t_0, \mathbf{T}]\} \leqq \delta_m(\Gamma_0), \quad \textit{if } n \geqq n(m).$$

Let $N = \{n(i): i = 1, 2, \cdots\}$, and define $\bar{\Gamma}$ as follows. $\bar{\Gamma} = (\bar{\Gamma}_m, \Pi_m)$ where $\Pi_m = \{t_0 < \tau_1 < \cdots < \tau_m = T\}$ is a uniform partition of $[t_0, T]$ into m subintervals of length $(T - t_0)/m$. Now if $m \notin N$, then let $\bar{\Gamma}_m$ be the mth-stage of the extremal strategy Γ_0. If $m \in N$, then $\bar{\Gamma}_{m,i} = u_m \restriction [t_0, \tau_i)$, for $i = 1, \cdots, k(m)$, where u_m is as defined above, and $k = k(m) \geqq 1$ is the smallest index such that

$$\max\{d(\bar{\varphi}_m(\tau), \mathbf{E}): \tau \in [\tau_{k-1}, \tau_k]\} > \delta_m(\Gamma_0), \tag{2.12}$$

and $\bar{\varphi}_m$ is the trajectory, with initial point (t_0, x_0), resulting from $\bar{\Gamma}_m$ versus some Δ_m on $[t_0, \tau_k]$. Note that if (2.12) never occurs or if $k = m$ then $\bar{\Gamma}_m$ has been defined completely. If $k < m$, then for every $j \geqq k+1$, define $\bar{\Gamma}_{m,j}$ extremally with respect to $[0, \mathbf{T}] \times \mathbf{E}$. Observe that,

$$d(\bar{\varphi}_m(\tau_k), \mathbf{E}) \leqq d(\bar{\varphi}_m(\tau_{k-1}), \mathbf{E}) + M(\mathbf{T} - t_0)/m \leqq \delta_m(\Gamma_0) + M(\mathbf{T} - t_0)/m,$$

where the second inequality follows from the definition of k. Therefore, if $m \in N$, using (2.10), we have that for all $\tau \in [\tau_k, T]$,

$$d(\bar{\varphi}_m(\tau), \mathbf{E}) \leqq \mathcal{K} d(\bar{\varphi}_m(\tau_k), \mathbf{E}) \leqq \mathcal{K}(\delta_m(\Gamma_0) + M(\mathbf{T} - t_0)/m).$$

By the definition of $\bar{\Gamma}$, if $m \notin N$, then $d(\bar{\varphi}_m(\tau), \mathbf{E}) \leqq \delta_m(\Gamma_0)$. Thus, $\bar{\Gamma} \in \mathbb{\Gamma}$. Furthermore, for $m \in N$, (u_m, v_m) is the mth-stage outcome of $\bar{\Gamma}$ and Δ_0. Therefore, $\varphi_m \in \Phi_m(\cdot, t_0, x_0, \bar{\Gamma}, \Delta_0)$ if $m \in N$. Since $\varphi_m \to \varphi$, we have $\varphi \in \Phi[\cdot, t_0, x_0, \bar{\Gamma}, \Delta_0]$. This proves the claim and the theorem. □

3. Regularity of the value. Although by Lemma 1.2, the value of the game considered in § 1 (i.e., with phase restrictions on only the maximizing player) is upper semicontinuous, it can easily fail to be continuous or even finite valued in general, as shown in the following example.

Example 1. Consider the game with dynamics

$$\dot{x} = y - z - \varepsilon$$

with $\varepsilon > 0$, $Y = [0, 1]$, $Z = [0, 1]$, $\mathbf{E} = \{x \in \mathbb{R}: x \geqq 0\}$, and $g(x) = x$. It is clear that $w(t, x) = -\infty$ for any initial point (t, x) with $0 \leqq x < (T - t)\varepsilon$, $\mathbf{T}$ = final time.

In this section we will discuss conditions which imply the finiteness, continuity, or Lipschitz continuity, of the value.

To avoid cumbersome notation, for our first two results (i.e., up to Theorem 3.3) we will assume that the phase set $\mathbf{E}$ is of a particularly simple type, namely:

$$\mathbf{E} = \{x \in \mathbb{R}^n: x^n \geqq 0\}. \tag{3.1}$$

These results extend easily to more general types of sets, as remarked below.

Clearly, the difficulty with Example 1, above, is that, for some initial values (t, x), the second player can always force the first player to violate his phase constraint. The

following condition will prevent such situations, ensuring the finiteness of the value.

H2. For any $(t, x) \in [0, \mathbf{T}] \times \partial \mathbf{E}$,

$$\max_{y} \min_{z} f^n(t, x, y, z) \geqq 0.$$

PROPOSITION 3.1. *If* **H1** *and* **H2** *holds then* $w(t, x)$ *is finite for all* $(t, x) \in [0, \mathbf{T}] \times \mathbf{E}$.

Proof. $w(t_0, x_0) = \sup_{\Gamma} \inf_{\Delta} P[t_0, x_0, \Gamma, \Delta]$ is finite valued if and only if there exists a strategy Γ such that for any $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma]$, $\varphi[t] \in \mathbf{E}$ for all $t \in [t_0, T]$; i.e., $\varphi^n[t] \geqq 0$, for all $t \in [t_0, \mathbf{T}]$. Let

$$\Omega(t, x) = \{y \in Y: \min_{z} f^n(t, x, y, z) \geqq 0\}.$$

Then, by **H2**, $\Omega(t, x) \neq \varnothing$, if $x^n = 0$. Consider the strategy $\Gamma = \{\Gamma_m, \Pi_m\}$ where the mth-stage partition $\Pi_m = \{t_0 < \tau_1 < \tau_2 < \cdots < \tau_m = \mathbf{T}\}$ is a uniform partition of $[t_0, \mathbf{T}]$ into m intervals. *Let* $\delta_m \equiv \|\Pi_m\| = (\mathbf{T} - t_0)/m$. Let Γ_m be defined as in [1, p. 189] using a positional strategy $U(t, x)$ satisfying $U(t, x) \in \Omega(t, x^1, \cdots, x^{n-1}, 0)$. Let $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma]$ with $\{\varphi_m(\cdot)\}$ as the sequence of mth-stage trajectories converging uniformly to φ.

Claim. $\varphi^n[t] \geqq 0$ for all $t \in [t_0, T]$. If not, then there exists a $\acute{t} \in [t_0, T]$ such that $\varphi^n[\acute{t}] < 0$. Let $\bar{t} = \inf\{t < \acute{t}: \varphi^n[\tau] < 0 \text{ for all } \tau \in (t, \acute{t}]\}$ and $\sigma = \acute{t} - \bar{t}$. We will arrive at a contradiction. By the definition of $\bar{t}$ and the continuity of φ, we have $\varphi^n[\bar{t}] = 0$, and $\varphi^n[t] < 0$ for all $t \in (\bar{t}, \bar{t} + \sigma]$. For each m, let $k = k(m)$ and $j = j(m)$ be the smallest integers satisfying

$$\text{(i)} \quad \bar{t} \in [t_0, \tau_k) \quad \text{and} \quad \text{(ii)} \quad \bar{t} + \sigma \in [\tau_j, \tau_{j+1})$$

Then,

$$\text{(3.2)} \qquad \text{(a)} \quad \tau_k \to \bar{t} \text{ and } \tau_j \to \bar{t} + \sigma \quad \text{as } m \to \infty.$$

$$\text{(b)} \quad \varphi^n_m(\tau_k) \to \varphi^n[\bar{t}] = 0 \quad \text{as } m \to \infty.$$

$$\text{(c)} \quad \varphi^n_m(t) < 0, \quad \forall t \in [\tau_k, \tau_j], \quad \text{for } m \text{ sufficiently large.}$$

Define $h(t, x, y) = \min\{f^n(t, x, y, z): z \in Z\}$. It follows from **H1** that

$$\text{(3.3)} \qquad \text{(a)} \quad |h(t, x, y) - h(t, \bar{x}, y)| \leqq K|x - \bar{x}|, \quad \forall t \in [0, T]; \ y \in Y; \quad x, \bar{x} \in \mathbb{R}^n,$$

$$\text{(b)} \quad h(t, x, U(t, x)) \geqq 0 \quad \text{if } x^n = 0,$$

where K = Lipschitz constant of f as a function of x. Statement (b) follows from the definition of $U(t, x)$. Let $R > 0$ be large enough so that $\max\{|\varphi_m(t)|: t \in [t_0, T]\} \leqq R$ for all m. Define $M = \max\{|f(t, x, y, z)|: t \in [0, T], |x| \leqq R, y \in Y, z \in Z\}$ and

$$\omega(r) = \sup\{|f^n(t, x, y, z) - f^n(t', x', y', z')|: |(t, x, y, z) - (t', x', y', z')| \leqq r, |x| \leqq R, |x'| \leqq R\},$$

then $\omega(r) \to 0$ as $r \to 0$. Now if $t \in [\tau_i, \tau_{i+1}]$, then

$$f^n(t, \varphi_m(t), y_i, v_m(t)) \geqq f^n(\tau_i, \varphi_m(\tau_i), y_i, v_m(t)) - \omega((M+1)\delta_m),$$

where $y_i \equiv U(\tau_i, \varphi_m(\tau_i))$. Hence

$$\text{(3.4)} \qquad \varphi^n_m(t) \geqq \varphi^n_m(\tau_i) - \delta_m \omega((M+1)\delta_m) + \int_{\tau_i}^{t} f^n(\tau_i, \varphi_m(\tau_i), y_i, v_m(\tau))\, d\tau.$$

By the definition of h, for almost every $\tau \in [\tau_i, t]$, we have

$$f^n(\tau_i, \varphi_m(\tau_i), y_i, v_m(\tau)) \geqq h(\tau_i, \varphi_m(\tau_i), y_i).$$

Let $\hat{\varphi}_m(t) \equiv (\varphi_m^1(t), \cdots, \varphi_m^{n-1}(t), 0)$, then by (3.3),

$$h(\tau_i, \varphi_m(\tau_i), y_i) \geqq h(\tau_i, \hat{\varphi}_m(\tau_i), y_i) - K|\varphi_m^n(\tau_i)| \geqq -K|\varphi_m^n(\tau_i)|.$$

Therefore, from (3.4), we get:

$$\varphi_m^n(t) \geqq \varphi_m^n(\tau_i) - \delta_m K|\varphi_m^n(\tau_i)| - \delta_m \omega((M+1)\delta_m), \quad \forall t \in [\tau_i, \tau_{i+1}).$$

Using (3.2.c), we obtain that if $k \leqq i \leqq j$ then

$$\varphi_m^n(t) \geqq \varphi_m^n(\tau_i)(1+\delta_m K) - \delta_m \omega((M+1)\delta_m). \tag{3.5}$$

It follows, using $\varphi_m^n(\tau_k) < 0$, that for any $t \in [\tau_k, \tau_j)$,

$$\varphi_m^n(t) \geqq \varphi_m^n(\tau_k)(1+\delta_m K)^{k-j} - \delta_m \omega((M+1)\delta_m)(1+\cdots+(1+\delta_m K)^{k-j-1}).$$

Since $k-j \leqq m$, we may replace $k-j$ by m (note that $\varphi_m^n(\tau_k) < 0$):

$$\varphi_m^n(t) \geqq \varphi_m^n(\tau_k)(1+\delta_k K)^m - \delta_m \omega((M+1)\delta_m)(1+\cdots+(1+\delta_m K)^m).$$

This gives, after summing the last expression on the right and regrouping, that for every $t \in [\tau_k, \tau_j]$,

$$\varphi_m^n(t) \geqq [\varphi_m^n(\tau_k) - \delta_m \omega((M+1)\delta_m)](1+\delta_m K)^m - \omega((M+1)\delta_m)/K.$$

Letting $m \to \infty$ and using (3.2.a), (3.2.b), the fact that $\delta_m = (T-t_0)/m$ and $\omega((M+1)\delta_m) \to 0$, we get $\varphi^n[t] \geqq 0$ for all $t \in [\bar{t}, \bar{t}+\sigma]$, contradicting the choice of $\bar{t}$ and σ, and proving the proposition. □

COROLLARY. *If E is as in* (3.1) *and* **H2** *holds, then for any $(t_0, x_0) \in [0, T] \times E$ the game with initial point (t_0, x_0) has a saddle point.*

Proof. To obtain the proof, combine Theorem 2.2 and Proposition 3.1. □

That condition **H2** is not sufficient in itself for the continuity of the value can be seen in the following example.

Example 2. Consider the game with dynamics:

$$\dot{x}^1 = 0, \qquad \dot{x}^2 = y - z, \qquad \dot{x}^3 = x^1 y,$$

control sets $Y = Z = [0, 1]$, $g(x) = x^2$, where $x = (x^1, x^2, x^3)$, and some $T > 0$ as the fixed final time. The phase set is $E = \{x \in \mathbb{R}^3 : x^3 \geqq 0\}$. Consider $x_0 \in \partial E$ (i.e., $x_0^3 = 0$) and some initial time $t_0 < T$. If $x_0^1 = 0$ then the optimal plays are $y^* \equiv 1$ and $z^* \equiv 1$ and $w(t_0, x_0) = x_0^2$. However if $x_0^1 < 0$ then the only admissible choice for player I is $y^* \equiv 0$. Hence, $w(t_0, x_0) = x_0^2 - (T - t_0)$. Therefore, on ∂E, w is discontinuous.

Remark. It is not difficult to find examples of games where w is discontinuous in the interior of $[0, T] \times E$. However, as will be shown in Theorem 3.3, if the value w is discontinuous on $[0, T] \times E$ then it is also discontinuous on $[0, T] \times \partial E$.

Let us also observe that the value in the following variant of the above example is continuous:

Example 3. In Example 2, let us change $\dot{x}^3$ to

$$\dot{x}^3 = x^1 y + \varepsilon$$

for some $\varepsilon > 0$. Consider the resulting game. For (t_0, x_0) (with $x_0 \in E$), the value is given by:

$$w(t_0, x_0) = \begin{cases} x_0^2 & \text{if } x_0^1 \geqq -\varepsilon \\ x_0^2 + ((\varepsilon/|x_0^1|) - 1)(T - \bar{t}) & \text{if } x_0^1 < -\varepsilon, \end{cases}$$

where $\bar{t} = \min(T, t_0 + x_0^3/(|x_0^1| - \varepsilon))$. Hence w is continuous.

Note. The existing results ([4], [5], and [8]) do not apply to this example because of their restrictive hypotheses.

We will show, in Theorem 3.1, that, in fact, the following strengthened versions of **H2** and **H1** ensure the continuity of the value:

H3. For every $t \in [0, T]$, $x \in \partial E$ (i.e., $x^n = 0$),

$$\max_y \min_z f^n(t, x, y, z) > 0.$$

H1′. Same as **H1** except for **H1**-(ii) which is changed to the following.

(ii) There exists a constant $K > 0$ such that

$$|f(t, x, y, z) - f(t', x', y, z)| \leqq K(|t - t'| + |x - x'|)$$

for all $t, t' \in [0, T]$, $x, x' \in E$, $y \in Y$, $z \in Z$.

We begin the discussion leading to Theorem 3.1 by introducing some of the parameters to be used in the proofs below.

Note 1. **H3** and our continuity assumptions on f imply that for any $R > 0$, there exist $\beta = \beta(R) > 0$ and $\varepsilon_0 = \varepsilon_0(R) > 0$ such that

$$\max_y \min_z f^n(t, x, y, z) \geqq \beta \tag{3.6}$$

for all $(t, x) \in [0, T] \times \{x \in \mathbb{R}^n : |x| \leqq R : |x| \leqq R, |x^n| \leqq \varepsilon_0\}$. We will denote by $\tilde{y}(t, x)$ any element of Y which attains the maximum in (3.6).

Note 2. Let $(\bar{t}, \bar{x}) \in [0, T] \times \{x \in \mathbb{R}^n : |x| \leqq R, |x^n| \leqq \varepsilon_0\}$, and let φ_v be a solution of

$$\dot{x} = f(t, x, \tilde{y}(\bar{t}, \bar{x}), v(t)), \quad \text{a.e. } t \in [\bar{t}, T], \qquad x(\bar{t}) = \bar{x},$$

where $v(\cdot)$ is some control function of player II. We will denote by $h_1 = h_1(R)$ a positive number such that

$$f^n(t, \varphi_v(t), \tilde{y}(\bar{t}, \bar{x}), v(t)) \geqq \beta/2, \quad \text{a.e. } t \in [\bar{t}, \bar{t} + h_1] \tag{3.7}$$

for any choice of $v(\cdot)$. That such an h_1 exists follows easily from (3.6) and the continuity assumptions on f.

Note 3. Let (t_0, x_0), (t_1, x_1) be given in $[0, T) \times X$, where X is a bounded subset of E. Let $s : [t_1, T] \mapsto [t_0, T]$ be defined by $s(\tau) = t_0 + (\tau - t_1)(T - t_0)/(T - t_1)$. Recall from [1, Lemma 6.4] that if **H1′** holds then there exists a constant $\tilde{K}$, independent of the initial points (t_i, x_i), such that if φ, φ' are solutions to

$$\dot{x} = f(t, x, u(t), v(t)), \quad \text{a.e. } t \in [t_1, T], \quad x(t_1) = x_1, \tag{3.8}$$

and

$$\dot{x} = f(t, x, u(s^{-1}(t)), v(s^{-1}(t))), \quad \text{a.e. } t \in [t_0, T], \qquad x(t_0) = x_0, \tag{3.9}$$

respectively, for some pair of controls $(u(\cdot), v(\cdot))$, then

$$\max \{|\bar{\varphi}(t) - \varphi(s(t))| : t \in [t_1, T]\} \leqq \tilde{K}(|t_0 - t_1| + |x_0 - x_1|). \tag{3.10}$$

We can now state and prove the main lemma leading to Theorem 3.1.

LEMMA 3.1. *Suppose* **H1′** *and* **H3** *hold. Then for every compact subset X of E, there exist $\mu > 0$, $\delta > 0$, and $C > 0$, depending on X, such that given a pair of points (t_0, x_0) and (t_1, x_1) in $[0, T) \times X$ with*

$$|t_0 - t_1| + |x_0 - x_1| < \delta, \tag{3.11}$$

there exists a strategy Γ^* on $[t_1, T]$ with the following property. If $\varphi^* \in \Phi[\cdot, t_1, x_1, \Gamma^*]$ and $t^* = \min(t_1 + \mu, T)$, then

$$\varphi^*[t] \in E, \quad \forall t \in [t_1, t^*]. \tag{3.12}$$

Moreover, there exists a motion $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma_0]$, *where* Γ_0 *is the extremal strategy defined with respect to* $C^+(w(t_0, x_0))$, *such that*

$$|\varphi^*[t] - \varphi[s(t)]| < C(|t_0 - t_1| + |x_0 - x_1|), \quad \forall t \in [t_1, t^*], \tag{3.13}$$

where $s(\cdot)$ *is as defined above.*

Proof. Let $R > 0$ be large enough so that all motions with initial points in $[0, T] \times X$ always remain in $B_R(0)$. This is possible using **H1'** and the compactness of X. Let β, h_1, ε_0, and $\tilde{K}$ be defined as in Notes 1-3 above with respect to this R. Let

$$\delta = \min(\varepsilon_0/2\tilde{K}, \beta h_1/4\tilde{K}). \tag{3.14}$$

To define Γ^*, note that, by the remark following Proposition 2.1, there exists an integer m^* such that any mth-stage trajectory resulting from Γ_0 satisfies

$$\max\{d(\varphi_m(t), E): t \in [t_0, T]\} < \varepsilon_0/2, \quad \text{if } m \geqq m^*. \tag{3.15}$$

We define Γ^*_m, the mth-stage of Γ^*, for $m \geqq m^*$. Let $\Pi_m = \{t_0 < \tau_1 < \cdots < \tau_m = T\}$ be a uniform partition of $[t_0, T]$, with norm $\|\Pi_m\| = (T - t_0)/m$. Set $\sigma_i = s^{-1}(\tau_i)$, $i = 1, \cdots, m$, and take Π^*_m, the mth-stage partition of Γ^*, to be $\{\sigma_0 < \sigma_1 < \cdots < \sigma_m\}$. Note that Π^*_m is also a uniform partition. Let Δ be any strategy of the second player. We will now describe how Γ^*_m chooses a control function on $[\sigma_i, \sigma_{i+1})$ playing against Δ_m.

To facilitate the description of Γ^*_m, let us denote by $u^*_{m,i}(\cdot)$, the choice of Γ^*_m on $[\sigma_i, \sigma_{i+1})$, and by $u^*_m(\cdot)$ the concatenation of $u^*_{m,i}$'s. Similarly, $v^*_{m,i}(\cdot)$ will be the choice of Δ_m on $[\sigma_i, \sigma_{i+1})$ and v^*_m their concatenation. Let u_m be the choice of $\Gamma_{0,m}$ on $[\tau_i, \tau_{i+1})$ playing against a control function v_m, which will be constructed using v^*_m. Once u^*_m and v^*_m are determined on an interval $[t_1, \sigma_i)$, we will denote by $\varphi^*_m(\cdot)$ the trajectory determined by (3.8) on $[t_1, \sigma_i]$ using u^*_m and v^*_m. Similarly, $\varphi_m(\cdot)$ will denote the trajectory on $[t_0, \tau_i]$ resulting from (u_m, v_m) with initial point (t_0, x_0).

We are now ready to define $\Gamma^*_{m,i}$, $i = 0, \cdots, m-1$. For $i = 0$, set $\Gamma^*_{m,0} = \Gamma_{0,m,0}(s(t))$, $t \in [t_1, \sigma_1)$. For $i > 0$, if $\varphi^*_m(\sigma_i) \in E$ then define

$$u^*_{m,i}(t) = u_{m,i}(s(t)), \qquad t \in [\sigma_i, \sigma_{i+1})$$

where $u_{m,i}$, defined on $[\tau_i, \tau_{i+1})$, is the choice of $\Gamma_{0,m}$ on $[\tau_i, \tau_{i+1})$ playing against v_m defined by

$$v_m(\tau) = v^*_m(s^{-1}(\tau)), \qquad \tau \in [t_0, \tau_{i-1}).$$

Let $k = k(m) > 1$ be the smallest integer such that $\varphi^*_m(\sigma_k) \notin E$. Define $h = (4\tilde{K}/\beta)(|t_0 - t_1| + |x_0 - x_1|)$. Note that $h < h_1$ by (3.11) and (3.14). Let $\nu = \nu(m)$ be determined as the smallest integer such that

$$[\sigma_k, \sigma_k + h) \subset \cup\{[\sigma_{k+i}, \sigma_{k+i+1}): i = 0, \cdots, \nu\}. \tag{3.16}$$

Note that $\sigma_k + h \in [\sigma_{k+\nu}, \sigma_{k+\nu+1}]$, and therefore

$$|\sigma_{k+\nu} - (\sigma_k + h)| \leqq \|\Pi^*_m\| \to 0, \quad \text{as } m \to \infty. \tag{3.17}$$

Now on the intervals $[\sigma_{k+i}, \sigma_{k+i+1})$, $i = 0, \cdots, \nu - 1$, define

$$u^*_{m,k+i}(t) \equiv \tilde{y}(\sigma_k, \varphi^*_m(\sigma_k)),$$

where $\tilde{y}(\cdot, \cdot)$ is as in Note 1 above. Observe that $\tilde{y}(\sigma_k, \varphi^*_m(\sigma_k))$ is defined since, by (3.15), $|\varphi^*_m(\sigma_k)| < \varepsilon_0$. Define $h_m \equiv \nu\|\Pi^*_m\|$. Then, since Π^*_m is a uniform partition, we have $\sigma_{k+\nu} = \sigma_k + h_m$. Let

$$u^*_{m,k+\nu}(t) = u_{m,k}(s(t - h_m)), \quad \text{for } t \in [\sigma_{k+\nu}, \sigma_{k+\nu+1}).$$

Note that $u_{m,k}(\cdot)$ is defined on $[\tau_k, \tau_{k+1})$ and depends only on the restrictions of u_m and v_m to $[t_0, \tau_k)$.

For $i \geqq \nu+1$, define $v_m(\cdot)$ on $[t_0, \tau_{k+i-\nu})$ by

$$v_m(\tau) = \begin{cases} v_m^*(s^{-1}(\tau)) & \text{if } \tau \in [t_0, \tau_k) \\ v_m^*(s^{-1}(\tau)+h_m) & \text{if } \tau \in [t_0, \tau_k) \\ z_0 & \text{if } \tau \in [\tau_k, \tau_{k+i-\nu}) \text{ and } s^{-1}(\tau)+h_m \geqq T, \end{cases}$$

where $z_0 \in Z$ is some arbitrarily chosen fixed element. Now, let $u_{m,k+i-\nu}$ be the outcome of $\Gamma_{0,m}$ against v_m, and for $i \geqq \nu+1$, define

$$u_{m,k+i}^*(t) = u_{m,k+i-\nu}(s(t-h_m)), \qquad t \in [\sigma_{k+i}, \sigma_{k+i+1}).$$

This completes the definition of Γ^*.

From the above constructions it follows that u_m^*, v_m^*, u_m, and v_m satisfy the following relationships:

$$u_m^*(t) = \begin{cases} u_m(s(t)) & t \in [t_1, \sigma_k) \\ \tilde{y}(\sigma_k, \varphi_m^*(\sigma_k)) & t \in [\sigma_k, \sigma_{k+\nu}) \\ u_m(s(t-h_m)) & t \in [\sigma_{k+\nu}, T], \end{cases} \tag{3.18}$$

and

$$v_m(\tau) = \begin{cases} v_m^*(s^{-1}(\tau)) & t \in [t_0, \tau_k) \\ v_m^*(s^{-1}(\tau)+hm) & t \in [\tau_k, s^{-1}(T-h_m)) \\ z_0 & t \in [s^{-1}(T-h_m), T]. \end{cases} \tag{3.19}$$

Let us also observe that since u_m is an outcome of Γ_0 (by its construction), φ_m is an mth-stage trajectory resulting from Γ_0.

For the analysis that follows we will need a third trajectory, $\bar{\varphi}_m(\cdot)$, defined on $[t_1, T]$ with controls $(\bar{u}_m, \bar{v}_m)$ and initial point (t_1, x_1), where

$$\text{(a)} \quad \bar{u}_m(t) = u_m(s(t)), \qquad t \in [t_1, T], \tag{3.20}$$

$$\text{(b)} \quad \bar{v}_m(t) = v_m(s(t)), \qquad t \in [t_1, T].$$

Observe that $\bar{\varphi}_m$ is an mth-stage trajectory resulting from $\Theta\Gamma_0$, where Θ is the map defined using $s(\cdot)$ as in [1] (or see Lemma 1.2 above).

We now verify that Γ^* has the desired properties. Let $\varphi^* \in \Phi[\cdot, t_1, x_1, \Gamma^*, \Delta]$, for some Δ. Let $\{\varphi_m^*\}$ be the sequence of mth-stage trajectories converging to φ^*. We may assume, without loss of generality, that $\varphi_m^*(t_1) = x_1$. Let $\varphi_m(\cdot) = \varphi_m(\cdot, t_0, x_0, u_m, v_m)$ and $\bar{\varphi}_m(\cdot) = \varphi_m(\cdot, t_1, x_1, \bar{u}_m, \bar{v}_m)$ where (u_m, v_m) and $(\bar{u}_m, \bar{v}_m)$ are related to (u_m^*, v_m^*) through (3.18)–(3.20). We claim that there exist constants $\mu > 0$ and $C > 0$, independent of φ^* and (t_i, x_i), $i = 0, 1$, such that

$$\text{(i)} \quad \max\{d(\varphi_m^*(t), E): t \in [t_1, t^*]\} \to 0, \quad \text{as } m \to \infty. \tag{3.21}$$

$$\text{(ii)} \quad \lim_{m\to\infty} |\varphi_m^*(t) - \varphi_m(s(t))| \leqq C(|t_0 - t_1| + |x_0 - x_1|).$$

Before proving (3.21) we finish the proof of the lemma using (3.21). It is clear that (3.21(i)) implies (3.12). As for (3.13), consider the sequence $\{\varphi_m\}$. Extracting a subsequence, if necessary, we may assume that, for some $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma_0, \Delta]$, $\varphi_m \to \varphi$, uniformly on $[t_0, T]$. Therefore, by (3.21(ii)), we have

$$|\varphi^*[t] - \varphi[s(t)]| < C(|t_0 - t_1| + |x_0 - x_1|).$$

Thus the lemma is proved once we establish (3.21). The proof of (3.21) is similar to that given by Soner, in a different context, in [11, Lemma 3.2].

Verification of (3.21(i)). Note that, by the definition of σ_k, $\varphi_m^*(t) \in E$ for all $t \in [t_1, \sigma_{k-1}]$, and all $m \geqq m^*$. Also, since $|\sigma_k - \sigma_{k-1}| = \|\Pi_m^*\| \to 0$ as $m \to \infty$, we get

$$d(\varphi_m^*(\sigma_k), E) \to 0, \quad \text{as } m \to \infty. \tag{3.22}$$

Recall that an $[\sigma_k, \sigma_{k+\nu}]$, $u_m^* \equiv \tilde{y}(\sigma_k, \varphi_m^*(\sigma_k))$. Hence, by (3.7),

$$f^n(t, \varphi_m^*(t), \tilde{y}(\sigma_k, \varphi_m^*(\sigma_k)), v_m^*(t)) \geqq \beta/2 > 0, \quad \forall t \in [\sigma_k, \sigma_{k+\nu}].$$

Therefore, $\max\{d(\varphi_m^*(t), E): t \in [\sigma_k, \sigma_{k+\nu}]\} = d(\varphi_m^*(\sigma_k), E)$. Hence, by (3.22), we always have

$$\max\{d(\varphi_m^*(t), E): t \in [t_1, \sigma_{k+\nu}]\} \to 0, \quad \text{as } m \to \infty. \tag{3.23}$$

Let $\mu > 0$ be such that

$$\gamma(\mu) \equiv K\mu\, e^{K\mu} + M(e^{K\mu} - 1) \leqq \beta/4, \tag{3.24}$$

where K is the Lipschitz constant of f (see **H1'**), and $M = \max\{|f(t, x, y, z)|: t \in [0, T], |x| \leqq R, y \in Y, z \in Z\}$. Now suppose $\sigma_{k+\nu} < t^*$ (otherwise we are done by (3.23)). Define $\tilde{t} = t^* - h_m$ (recall $h_m = \nu\|\Pi\|_m^*\|$), and $\psi(\cdot)$ by

$$\psi(t) = (\varphi_m^*)^n(t + h_m) - (\bar{\varphi}_m)^n(t), \quad \text{for } t \in [\sigma_k, \tilde{t}],$$

where $\bar{\varphi}_m$ is related to φ_m^* as explained before. Then, for $t \in [\sigma_k, \tilde{t}]$,

$$\begin{aligned} \psi(t) = \psi(\sigma_k) + \int_{\sigma_k}^{t} [&f^n(s + h_m, \varphi_m^*(s + h_m), u_m^*(s + h_m), v_m^*(s + h_m)) \\ &- f^n(s, \bar{\varphi}_m(s), \bar{u}_m(s), \bar{v}_m(s))]\, ds. \end{aligned} \tag{3.25}$$

Let I denote the integral on the right, then, by (3.18) and (3.19),

$$I = \int_{\sigma_k}^{t} [f^n(s + h_m, \varphi_m^*(s + h_m), \bar{u}_m(s), \bar{v}_m(s)) - f^n(s, \bar{\varphi}_m(s), \bar{u}_m(s), \bar{v}_m(s))]\, ds.$$

By the Lipschitz continuity of f,

$$|I| \leqq \int_{\sigma_k}^{t} K[h_m + |\varphi_m^*(s + h) - \bar{\varphi}_m(s)|]\, ds, \qquad t \in [\sigma_k, \tilde{t}]. \tag{3.26}$$

In a similar way, for $s \in [\sigma_k, \tilde{t}]$,

$$|\varphi_m^*(s + h_m) - \bar{\varphi}_m(s)| \leqq |\varphi_m^*(\sigma_k + h_m) - \bar{\varphi}_m(\sigma_k)| + \int_{\sigma_k}^{s} K[h_m + |\varphi_m^*(\tau + h_m) - \bar{\varphi}_m(\tau)|]\, ds.$$

Using Gronwall's lemma, we get, for $s \in [\sigma_k, \tilde{t}]$,

$$|\varphi_m^*(s + h_m) - \bar{\varphi}_m(s)| \leqq [|\varphi_m^*(\sigma_k + h_m) - \bar{\varphi}_m(\sigma_k)| + Kh_m\mu]\, e^{K(s - \sigma_k)}. \tag{3.27}$$

Also, since $\varphi_m^*(\sigma_k) = \bar{\varphi}_m(\sigma_k)$, we have

$$|\varphi_m^*(\sigma_k + h_m) - \bar{\varphi}_m(\sigma_k)| \leqq \int_{\sigma_k + h_m}^{\sigma_k} |f(s, \varphi_m^*(s), u_m^*(s), v_m^*(s))|\, ds \leqq Mh_m.$$

Together with (3.27) we obtain

$$|\varphi_m^*(s + h_m) - \bar{\varphi}_m(s)| \leqq (M + K\mu)h_m\, e^{K(s - \sigma_k)}, \qquad s \in [\sigma_k, \tilde{t}]. \tag{3.28}$$

Substituting in (3.26) and integrating we get

$$|I| \leqq h_m(K\mu\, e^{K\mu} + M(e^{K\mu} - 1)) \equiv h_m\gamma(\mu), \tag{3.29}$$

(where $\gamma(\cdot)$ was defined in (3.24)). Thus, from (3.25) and (3.29), we obtain

$$\psi(t) \geqq \psi(\sigma_k) - h_m \gamma(\mu), \qquad t \in [\sigma_k, \tilde{t}]. \tag{3.30}$$

On the other hand,

$$\psi(\sigma_k) = \varphi_m^{*n}(\sigma_k + h_m) - \bar{\varphi}_m^n(\sigma_k) = \int_{\sigma_k}^{\sigma_k + h_m} f^n(s, \varphi_m^*(s), u_m^*(s), v_m^*(s))\, ds. \tag{3.31}$$

Recall that $\sigma_k + h \in [\sigma_{k+\nu}, \sigma_{k+\nu+1}]$. Since, by definition, $\sigma_k + h_m = \sigma_{k+\nu}$, we have, using (3.17),

$$|h_m - h| \leqq \|\Pi_m^*\| \to 0, \quad \text{as } m \to \infty.$$

Since $h < h_1$, there exists an m_1 such that $h_m < h_1$ if $m > m_1$. Hence, by (3.7) and (3.31),

$$\psi(\sigma_k) \geqq \beta h_m / 2, \quad \text{for } m > m_1.$$

Substituting into (3.30), we obtain

$$\psi(t) \geqq h_m[\beta/2 - \gamma(\mu)], \qquad t \in [\sigma_k, \tilde{t}], \qquad m > m_1. \tag{3.32}$$

By the choice of μ (see (3.20)),

$$\psi(t) \geqq h_m \beta / 4. \tag{3.33}$$

Next, noting the relationship between φ_m and $\bar{\varphi}_m$, we have from (3.10),

$$\max\{|\bar{\varphi}_m(t) - \varphi_m(s(t))| : t \in [t_1, T]\} \leqq \tilde{K}(|t_0 - t_1| + |x_0 - x_1|). \tag{3.34}$$

Hence if $\rho_m = \max\{d(\varphi_m(t), E) : t \in [t_0, T]\}$, then, using (3.34), we get

$$\max\{d(\bar{\varphi}_m(t), E) : t \in [t_0, \tilde{t}]\} \leqq \rho_m + \tilde{K}(|t_0 - t_1| + |x_0 - x_1|). \tag{3.35}$$

Observe that $\varphi_m(\cdot)$ is an mth-stage trajectory resulting from Γ_0. By Lemma 1.5, $\rho_m \to 0$, as $m \to \infty$. From (3.33) and the definition of $\psi(\cdot)$, we have

$$\begin{aligned} \varphi_m^{*n}(t + h_m) &\geqq \beta h_m / 4 + \bar{\varphi}_m^n(t) \\ &\geqq \beta h_m / 4 - \max\{d(\bar{\varphi}_m(t), E) : t \in [t_0, \tilde{t}]\}. \end{aligned}$$

Taking the minimum of the left side on $[\sigma_k, \tilde{t}]$, and using (3.35), we get

$$\min\{\varphi_m^{*n}(t + h_m) : t \in [\sigma_k, \tilde{t}]\} \geqq \beta h_m / 4 - \tilde{K}(|t_0 - t_1| + |x_0 - x_1|) - \rho_m.$$

Since $h_m \to h$ $(= 4\tilde{K}(|t_0 - t_1| + |x_0 - x_1|)/\beta)$ and $\rho_m \to 0$ as $m \to \infty$, we conclude

$$\lim_{m \to \infty} \min\{\varphi_m^{*n}(t + h_m) : t \in [\sigma_k, \bar{t}]\} \geqq 0.$$

This is the desired conclusion since $\sigma_k + h_m = \sigma_{k+\nu}$ and $\tilde{t} + h_m = t^*$.

Verification of (3.21(ii)). By the triangle inequality,

$$|\varphi_m^*(t) - \varphi_m(s(t))| \leqq |\varphi_m^*(t) - \bar{\varphi}_m(t)| + |\bar{\varphi}_m(t) - \varphi_m(s(t))|. \tag{3.36}$$

We already have an estimate of the second term on the right, namely (3.34). Let us consider the first term.

Case 1. $\sigma_{k+\nu} \leqq t^*$. Recall that for any $t \in [t_1, \sigma_k]$, $\varphi_m^*(t) = \bar{\varphi}_m(t)$. Now, if $t \in (\sigma_k, \tilde{t}]$, then

$$\begin{aligned} |\varphi_m^*(t) - \bar{\varphi}_m(t)| &\leqq |\varphi_m^*(t + h_m) - \bar{\varphi}_m(t)| + |\varphi_m^*(t + h_m) - \varphi_m^*(t)| \\ &\leqq (M + k\mu) h_m e^{K(t - \sigma_k)} + M h_m \qquad \text{(using (3.28))} \\ &\leqq h_m[(M + k\mu) e^{K\mu} + M]. \end{aligned}$$

If $t \in [\tilde{t}, t^*]$, then

$$|\varphi_m^*(t) - \bar{\varphi}_m(t)| \leqq |\varphi_m^*(t-h_m) - \varphi_m^*(t)| + |\varphi_m^*(t-h_m) - \bar{\varphi}_m(t-h_m)| + |\bar{\varphi}_m(t-h_m) - \bar{\varphi}_m(t)|$$

$$\leqq \int_{t-h_m}^{t} |f(\tau, \varphi_m^*(\tau), u_m^*(\tau), v_m^*(\tau))|\, d\tau + h_m[(M+k\mu)\, e^{K\mu} + M] + \int_{t-h_m}^{t} |f(\tau, \bar{\varphi}_m(\tau), \bar{u}_m(\tau), \bar{v}(\tau))|\, d\tau$$

$$\leqq 2h_m M + h_m[(M+k\mu)\, e^{K\mu} + M] = h_m[3M + (M+k\mu)\, e^{K\mu}].$$

Case 2. $\sigma_{k+\nu} > t^*$. Here, since $\varphi_m^*(\sigma_k) = \bar{\varphi}_m(\sigma_k)$, we have, for $t \in [\sigma_k, t^*]$,

$$|\varphi_m^*(t) - \bar{\varphi}_m(t)| \leqq \int_{\sigma_k}^{t} |f(\tau, \varphi_m^*(\tau), u_m^*(\tau), v_m^*(\tau)) - f(\tau, \bar{\varphi}_m(\tau), \bar{u}_m(\tau), \bar{v}(\tau))|\, d\tau$$

$$\leqq 2M(t-\sigma_k) \leqq 2Mh_m,$$

where the last inequality holds since $(t-\sigma_k) \leqq (t^*-\sigma_k) \leqq \sigma_{k-\nu} - \sigma_k \leqq h_m$. Hence in both cases

$$|\varphi_m^*(t) - \bar{\varphi}_m(t)| \leqq h_m[3M + (M+K\mu)\, e^{K\mu}], \quad \forall t \in [\sigma_k, t^*].$$

Now, using this and the fact that $h_m \to h = (4/\beta)\tilde{K}(|t_0 - t_1| + |x_0 - x_1|)$, we have

$$\lim_{m\to\infty} |\varphi_m^*(t) - \bar{\varphi}_m(t)| \leqq \tilde{C}(|t_0 - t_1| + |x_0 - x_1|), \quad \forall t \geqq \sigma_k, \tag{3.37}$$

where $\tilde{C} = (4/\beta)\tilde{K}((M+K\mu)\, e^{K\mu} + M)$.

Combining (3.37), (3.34), and using (3.36), we obtain (3.21(ii)) with $C = \tilde{C} + \tilde{K}$. Note that C depends only on $R > 0$ and f. This completes the proof of Lemma 3.1. □

Before stating Theorem 3.1, we make some technical notes which will be used in the proof of the theorem.

Note i. For the assertion of Lemma 3.1 to be applicable to a pair of points (t_0, x_0), (t_1, x_1), we need $|t_0 - t_1| + |x_0 - x_1| < \min(\beta h_1/4\tilde{K}, \varepsilon_0/2\tilde{K})$. It would be convenient, below, to also assume $|t_0 - t_1| < \mu/2$. Since μ depends only on $R > 0$ and f, this causes no problem. Hence, assume for Lemma 3.1 that

$$|t_0 - t_1| + |x_0 - x_1| < \delta^* \triangleq \min(\beta h_1/4\tilde{K}, \varepsilon_0/2\tilde{K}, \mu/2). \tag{3.38}$$

Note ii. Assume (3.38) holds. Note that if $t_0, t_1 \in [T-\mu, T]$ then

$$s(t_1 + \mu) \geqq T, \tag{3.39}$$

since $t_1 + \mu \geqq T$, $s(T) = T$, and $s(\cdot)$ is nondecreasing on $[t_1, T]$. If $t_0, t_1 \in [0, T-\mu]$, then

$$s(t_1 + \mu) \geqq t_0 + \mu/2. \tag{3.40}$$

To see this, recall that $s(t) = m(t - t_1) + t_0$, with $m = (T - t_0)/(T - t_1)$. Now if $t_1 \geqq t_0$, then $m \geqq 1$ and $s(t_1 + \mu) \geqq t_0 + \mu$. If $t_1 < t_0$, then

$$0 < t_0 - t_1 < \mu/2 < (T - t_1)/2.$$

Now $2(t_0 - t_1) < T - t_1$ is equivalent to $m \geqq 1/2$. Hence (3.40) holds.

Note iii. Let $\alpha(t) = t - s(t) = (1-m)t + (mt_1 - t_0)$, $t \in [t_1, T]$. Since $\alpha(\cdot)$ is a linear function, $\max\{|\alpha(t)|: t \in [t_1, T]\}$ occurs either at t_1 or at T. But $\alpha(T) = 0$. Hence, $|\alpha(t_1)| \geqq |\alpha(t)|$, for all $t \in [t_1, T]$. That is,

$$|t_1 - t_0| \geqq |t - s(t)| \quad \forall t \in [t_1, T]. \tag{3.41}$$

Note iv. In the proof of Theorem 3.1, as well as some other results in the later sections, we will use inductive arguments. The following device will be useful in these arguments. Let $R>0$ and denote by $S(R)$ the set of all initial points $(t_0, x_0)\in[0, T]\times E$ such that any solution $\varphi(\cdot)$ of the dynamics (0.1) with $\varphi(t_0)=x_0$ satisfies $|\varphi(t)|\leqq R$ for all $t\in[t_0, T]$. It follows from **H1** that given any $(t_0, x_0)\in[0, T]\times E$, there exists $R_0>0$ such that $(t_0, x_0)\in S(R_0)$. Hence $[0, T]\times E=\cup\{S(R): R>0\}$. Furthermore the following two properties hold.

P1. Let $(t_0, x_0)\in S(R_0)$. Then for every motion $\varphi[\cdot, t_0, x_0, \Gamma, \Delta]$, resulting from some pair of strategies Γ and Δ, and for every $t\in[t_0, T]$, we have

$$(t, \varphi[t, t_0, x_0, \Gamma, \Delta])\in S(R).$$

P2. $S(R)$ is closed (hence compact).

We prove **P1**. Property **P2** can be proved along similar lines.

Proof of **P1**. Let $t_1\in[t_0, T]$ and $\varphi\in\Phi[\cdot, t_0, x_0, \Gamma, \Delta]$ for some Γ and Δ. Set $x_1=\varphi[t_1]$. Let $\{\varphi_m(\cdot, t_0, x_{0m}, u_m(\cdot), v_m(\cdot))\}$ (we will write $\varphi_m(\cdot)$ for convenience) be a sequence of mth-stage trajectories converging to φ. We need to show that if, for some pair of controls $(u(\cdot), v(\cdot))$, $\eta(\cdot)$ satisfies

$$\dot\eta(t)=f(t, \eta(t), u(t), v(t)), \quad \text{a.e. } t\in[t_1, T], \eta(t_1)=x_1,$$

then $|\eta(t)|\leqq R$. Define $\tilde u_m$ and $\tilde v_m$ by

$$\tilde u_m(t)=\begin{cases} u_m(t) & \text{if } t\in[t_0, t_1)\\ u(t) & \text{if } t\in[t_1, T]\end{cases} \quad\text{and}\quad \tilde v_m(t)=\begin{cases} v_m(t) & \text{if } t\in[t_0, t_1)\\ v(t) & \text{if } t\in[t_1, T],\end{cases}$$

and let $\tilde\varphi_m(\cdot)=\tilde\varphi_m(\cdot, t_0, x_0, \tilde u_m(\cdot), \tilde v_m(\cdot))$. Since $(t_0, x_0)\in S(R)$, $|\tilde\varphi_m(t)|\leqq R$ for all $t\in[t_0, T]$. By the continuous dependence of solutions of differential equations on their initial values, we have

$$\eta(t)=\lim_{m\to\infty}\tilde\varphi_m(t), \quad \text{for } t\in[t_1, T].$$

Hence $|\eta(t)|\leqq R$ for all $t\in[t_1, T]$. □

In what follows we will use the following notation. If Ω is a subset of $\mathbb{R}\times\mathbb{R}^n$, let $\Omega_t\equiv\{(s, x)\in\Omega: s=t\}$, and $\Omega_{[\tau,t]}\equiv\{(s, x)\in\Omega: s\in[\tau, t]\}$. We now state and prove the following theorem.

THEOREM 3.1. *If* **H1′** *and* **H3** *hold then* $w(\cdot,\cdot)$ *is continuous (or locally Lipschitz continuous) on* $[0, T]\times E$ *if* $g(\cdot)$ *is.*

Proof. We will prove that if g is locally Lipschitz then so is w; that is, for every compact set $X\subset E$, there exists a constant C^* such that for every (t_0, x_0), $(t_1, x_1)\in[0, T]\times X$,

$$|w(t_0, x_0)-w(t_1, x_1)|\leqq C^*(|t_0-t_1|+|x_0-x_1|).$$

The modifications needed for the other statement are straightforward.

Let $R>0$ be such that $S(R)$ contains $[0, T]\times X$. Since by **P2**, above, $S(R)$ is compact, we obtain from Lemma 3.1 the constants $\mu>0$, $C>0$, $\tilde K>0$, $h_1>0$, $\delta^*>0$, etc. Now, for some positive integer σ,

$$S(R)\subset\bigcup_{k=1}^{\sigma} S(R)_{[T-k\mu/2, T]}.$$

We will show, by induction, that for each $k=2, \cdots, \sigma$ there exist constants $\bar\delta_k$, $\bar C_k$ such that for every (t_0, x_0), and (t_1, x_1) in $S(R)_{[T-k\mu/2, T]}$, we have

$$|t_0-t_1|+|x_0-x_1|<\bar\delta_k\Rightarrow|w(t_0, x_0)-w(t_1, x_1)|\leqq\bar C_k(|t_0-t_1|+|x_0-x_1|). \tag{3.42}$$

Then, since $S(R) \supset [0, T] \times X$, it follows that (3.42), with $k = \sigma$, holds for every (t_0, x_0), $(t_1, x_1) \in [0, T] \times X$. We then show that this gives the desired conclusion. We now proceed with the proof of (3.42).

Case $k = 2$ (i.e., $t_i \in [T - \mu, T]$). If $t_0 = T$, then we let Γ^* be the extremal strategy with respect to $C^+(w(t_1, x_1))$. Note that for all $\varphi \in \Phi[\cdot, t_1, x_1, \Gamma^*]$, $\varphi[t] \in E$ for all $t \in [t_1, T]$ and since φ is a Lipschitz function with constant M—where M is as defined following (3.2)—we obtain,

$$|\varphi[T] - x_0| \leqq |\varphi[T] - \varphi[t_1]| + |x_1 - x_0| \leqq M\{|t_0 - t_1| + |x_1 - x_0|\},$$

where we assumed, without loss of generality, that $M \geqq 1$ and used $T = t_0$ in the last inequality. Therefore, by the Lipschitz continuity of $g(\cdot)$, we obtain

$$w(t_1, x_1) \geqq w(t_0, x_0) - C_g M\{|t_0 - t_1| + |x_1 - x_0|\},$$

where C_g is the Lipschitz constant of g. The case $t_1 = T$ leads to the same inequality in a similar way using Γ_0, the strategy extremal to $C^+(w(t_0, x_0))$, in place of Γ^*, and noting that $w(t_1, x_1) = g(x_1)$.

Now let us assume $t_0, t_1 \in [T - \mu, T)$. Take $\bar{\delta}_2 = \delta^*$ (see (3.38)). Let Δ be any strategy of the second player on $[t_1, T]$ and Γ^* be as in Lemma 3.1. If $\varphi^* \in \Phi[\cdot, t_1, x_1, \Gamma^*, \Delta]$ then, since $t_1 + \mu \geqq T$, by that lemma we have

$$\varphi^*[t] \in E \quad \text{for every } t \in [t_1, T]. \tag{3.43}$$

Moreover, there exists $\varphi \in \Phi[\cdot, t_1, x_1, \Gamma_0]$, where Γ_0 is as defined above, such that

$$|\varphi^*[t] - \varphi[s(t)]| \leqq C(|t_0 - t_1| + |x_0 - x_1|) \quad \forall t \in [t_1, T]. \tag{3.44}$$

Since $s(T) = T$, and Γ_0 is extremal with respect to $C^+(w(t_0, x_0))$, we have $g(\varphi[s(T)]) \geqq w(t_0, x_0)$. Using (3.43), (3.44) and the Lipschitz continuity of g, we get

$$V(\varphi^*) = g(\varphi^*[T]) \geqq w(t_0, x_0) - \bar{C}_2(|t_0 - t_1| + |x_0 - x_1|) \tag{3.45}$$

with $\bar{C}_2 = C_g C$. Hence, taking sup over all strategies Γ, we obtain

$$\sup_{\Gamma} P[t_1, x_1, \Gamma, \Delta] \geqq w(t_0, x_0) - \bar{C}_2(|t_0 - t_1| + |x_0 - x_1|).$$

Since Δ was chosen arbitrarily, taking the infimum over all Δ on the left, we get

$$w(t_1, x_1) \geqq w(t_0, x_0) - \bar{C}_2(|t_0 - t_1| + |x_0 - x_1|).$$

We may assume that $\bar{C}_2 \geqq C_g M$ so that the above inequality holds for all $t_0, t_1 \in [T - \mu, T]$. Switching the roles of (t_0, x_0) and (t_1, x_1), we obtain the same inequality with (t_0, x_0) and (t_1, x_1) interchanged. This proves (3.42) in this case.

Now assume (3.42) holds for k, and consider the case for $k+1$; i.e., $t_i \in [T - (k+1)\mu/2, T]$, $i = 0, 1$. We distinguish two possibilities:

First. t_0, t_1 are both in $[T - (k+1)\mu/2, T - k\mu/2]$. Then we take

$$\bar{\delta}_{k+1} = \min(\delta^*, \bar{\delta}_k/2C, \bar{\delta}_k/2)$$

and suppose $|t_0 - t_1| + |x_0 - x_1| < \bar{\delta}_{k+1}$. As before, let Δ be an arbitrary strategy of the second player on $[t_1, T]$ and $\varphi^* \in \Phi[\cdot, t_1, x_1, \Gamma^*, \Delta]$. Then

$$\varphi^*[t] \in E \quad \text{for every } t \in [t_1, t_1 + \mu], \tag{3.46}$$

and there is a motion $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma_0]$ such that

$$|\varphi^*[t] - \varphi[s(t)]| \leqq C(|t_0 - t_1| + |x_0 - x_1|) < \bar{\delta}_k/2, \quad \forall t \in [t_1, t_1 + \mu]. \tag{3.47}$$

Now note that $s(t_1+\mu)\in[T-k\mu/2, T]$, by (3.40). Clearly $t_1+\mu\in[T-k\mu/2, T]$, also. Moreover, it follows from (3.41), and the definition of $\bar{\delta}_{k+1}$, that

$$|t_1+\mu-s(t_1+\mu)|\leqq|t_1-t_0|<\bar{\delta}_{k+1}\leqq\bar{\delta}_k/2. \tag{3.48}$$

Thus, by (3.47) and (3.48), we have

$$|t_1+\mu-s(t_1+\mu)|+|\varphi^*[t_1+\mu]-\varphi[s(t_1+\mu)]|<\bar{\delta}_k.$$

Using our induction hypothesis, and property **P1** of $S(R)$, we have:

$$\begin{aligned}
&|w(t_1+\mu, \varphi^*[t_1+\mu])-w(s(t_1+\mu), \varphi[s(t_1+\mu)])|\\
&\qquad\leqq\bar{C}_k(|t_1+\mu-s(t_1+\mu)|+|\varphi^*[t_1+\mu]-\varphi[s(t_1+\mu)]|)\\
&\qquad\leqq\bar{C}_k(|t_0-t_1|+C(|t_0-t_1|+|x_0-x_1|)) \quad \text{by (3.47)}\\
&\qquad\leqq\bar{C}_k(C+1)(|t_0-t_1|+|x_0-x_1|).
\end{aligned} \tag{3.49}$$

Let Γ_e be defined on $[t_1+\mu, T]$ extremally with respect to $C^+(w(t_1+\mu, \varphi^*[t_1+\mu]))$. Define $\bar{\Gamma}$ as the concatenation of Γ^* on $[t_1, t_1+\mu]$ with Γ_e on $[t_1+\mu, T]$. Then there exists a motion $\bar{\varphi}\in\Phi[\cdot, t_1, x_1, \bar{\Gamma}, \Delta]$ such that

$$\varphi[t]=\varphi^*[t]\in E \quad \forall t\in[t_1, t_1+\mu]$$

and

$$\bar{\varphi}\restriction[t_1+\mu, T]\in\Phi[\cdot, t_1+\mu, \varphi^*[t_1+\mu], \Gamma_e].$$

Therefore, by the properties of extremal strategies,

$$V(\bar{\varphi})=g(\bar{\varphi}[T])\geqq w(t_1+\mu, \varphi^*[t_1+\mu]).$$

Using (3.49) and the fact that $w(s(t_1+\mu), \varphi[s(t_1+\mu)])\geqq w(t_0, x_0)$, we obtain

$$V(\bar{\varphi})\geqq w(t_0, x_0)-C_k(C+1)(|t_0-t_1|+|x_0-x_1|).$$

Taking the sup over all Γ's and the inf over all Δ's, as in case $k=2$, we get $w(t_1, x_2)\geqq w(t_0, x_0)-C_k(C+1)(|t_0-t_1|+|x_0-x_1|)$. Switching the roles of (t_0, x_0) and (t_1, x_1), we obtain

$$|w(t_0, x_0)-w(t_1, x_1)|\leqq C_k(C+1)(t_0-t_1|+|x_0-x_1|).$$

Second. One of t_0, t_1, say t_1, is in $(T-k\mu/2, T]$. Let us define $(t_2, x_2)\equiv(T-k\mu/2, x_1)$. Then, with $\bar{\delta}_{k+1}$ as before,

$$|t_2-t_1|+|x_2-x_1|\leqq|t_2-t_1|\leqq|t_1-t_0|\leqq\bar{\delta}_k/2,$$

and

$$|t_2-t_0|+|x_2-x_0|\leqq|t_0-t_1|+|x_0-x_1|\leqq\bar{\delta}_{k+1}.$$

Therefore, using the first case and our induction hypothesis,

$$\begin{aligned}
|w(t_0, x_0)-w(t_1, x_1)|&\leqq|w(t_0, x_0)-w(t_2, x_2)|+|w(t_2, x_2)-w(t_1, x_1)|\\
&\leqq\bar{C}_k(C+1)(|t_0-t_1|+|x_0-x_1|)+\bar{C}_k(|t_1-t_0|)\\
&\leqq\bar{C}_{k+1}(|t_0-t_1|+|x_0-x_1|),
\end{aligned}$$

with $\bar{C}_{k+1}=\bar{C}_k(C+2)$. Hence in either case, (3.42) is satisfied with $\bar{C}_{k+1}$ and $\bar{\delta}_{k+1}$.

This concludes our induction, proving (3.42). It remains to show that the conclusion of the theorem follows from (3.42). Note that it follows from (3.42) that $w(\cdot, \cdot)$ is continuous on $[0, T]\times X\subset S(R)$. In particular, there exists a constant $L>0$ such that

$$|w(t, x)|\leqq L \quad \forall(t, x)=[0, T]\times X.$$

Let $C^*=\max(\bar{C}_\sigma, 2L/\bar{\delta}_\sigma)$. Then given a pair of points (t_0, x_0) and (t_1, x_1) in $[0, T]\times X$,

$$|t_0-t_1|+|x_0-x_1|<\bar{\delta}_\sigma\Rightarrow|w(t_0, x_0)-w(t_1, x_1)|\leqq C^*(|t_0-t_1|+|x_0-x_1|),$$

(using (3.42)) and

$$|t_0-t_1|+|x_0-x_1|\geqq\bar{\delta}_\sigma \Rightarrow |w(t_0,x_0)-w(t_1,x_1)|\leqq 2L\leqq C^*(|t_0-t_1|+|x_0-x_1|).$$

That is, $w(\cdot,\cdot)$ is Lipschitz on $[0,T]\times X$ with constant C^*. The proof of the theorem is now complete. □

Condition **H3** is not necessary for the continuity of the value. This can be seen from the following example:

Example 4. Consider Example 2, again:

$$\dot{x}^1=0$$
$$\dot{x}^2=y-z$$
$$\dot{x}^3=x^1y,$$

$Y=Z=[0,1]$, $E=\{x: x^3\geqq 0\}$. Here, however, we change the payoff function, $g(\cdot)$, to $g(x)=-x^2$. It can easily be verified that for any $(t_0,x_0)\in[0,t]\times E$, $w(t_0,x_0)=-x_0^2$. Thus w is continuous on $[0,T]\times E$.

Note that **H3** is not satisfied here but the optimal play, $y^*\equiv 0$, automatically gives $\dot{x}^3\geqq 0$. In some examples, such as the above, it can be directly verified that the extremal strategies automatically respect the given state constraints. Condition **H4**, below, says just this. We will show in Theorem 3.2 that **H1** and **H4** ensure that the value is continuous (or Lipschitz continuous) if g is continuous (Lipschitz continuous).

H4. (i) $\exists x\in E$ such that $w(0,x)>-\infty$,

(ii) If $t\in[0,T]$, $x\notin E$ and $\alpha\in\mathbb{R}$ are such that $U^\alpha(t,x)$, the extremal strategy with respect to $C^+(\alpha)$, is defined then

$$\min_z f^n(t,x,y^*(t,x),z)\geqq 0,$$

where $y^*(t,x)$ is the outcome of $U^\alpha(t,x)$.

Remarks. 1. By Proposition 3.1, **H2** implies **H4**-(i).

2. **H4** implies **H2**; in particular, $w(t,x)\in\mathbb{R}$, for all (t,x) in $[0,T]\times E$. To see this, note that, by **H4**-(i), there exists an $x_0\in E$ such that $w(0,x_0)\in\mathbb{R}$. Let $\alpha=w(0,x_0)$. Then $U^\alpha(t,x)$ is defined for all (t,x) in $[0,T]\times E$ since for all $t\in[0,T]$, $C^+(\alpha)\cap\{(t,x): x\in R^n\}\neq\varnothing$. Now given (t,x) with $x^n=0$, let $x_\kappa=(x^1,x^2,\cdots,x^{n-1},1/k)$.Then, using **H4**-(ii),

$$\max_y\min_z f^n(t,x_\kappa,y,z)\geqq\min_z f^n(t,x_\kappa,y^*(t,x_\kappa),z)\geqq 0.$$

Since $(t,x)\to\max_y\min_z f^n(t,x,y,z)$ is continuous, **H2** follows.

3. **H4**-(ii) is always satisfied if the following conditions (a), (b) hold:

(a) $Y=Y_1\times Y_2$, and $Z=Z_1\times Z_2$, for some compact sets Y_i and Z_i, such that

$$f^i(t,x,y,z)=f^i(t,x,y_1,z_1),\quad \text{for } i=1,\cdots,n-1,$$

and

$$f^n(t,x,y,z)=f^n(t,x,y_1,y_2,z_2);$$

i.e., f^n has controls distinct from the controls of $f^1,\cdots,f^{n-1}$.

(b) For every (t,x,y_1,z_2), if $x\notin E$ then $\exists y_2\in Y_2$ such that $f^n(t,x,y_2,z_2)\geqq 0$.

Proof. Suppose $y^*(t,x)=(y_1^*,y_2^*)$ is the outcome of an extremal strategy $U^\alpha(t,x)$ at (t,x), $x\notin E$. Then, by the definition of extremal strategies, there exist some $z^*=(z_1^*,z_2^*)\in Z$ and $w\in C^+(\alpha)$ such that for all $z\in Z$, $y\in Y$,

$$\langle s,f(t,x,y^*,z)\rangle\geqq\langle s,f(t,x,y^*,z^*)\rangle\geqq\langle s,f(t,x,y,z^*)\rangle,$$

where $s = w - x$. Note that $w^n \geqq 0$, since $C^+(\alpha) \subset [0, T] \times E$, and $x^n < 0$. Hence, $s^n > 0$. Let $\hat{v}$ denote the first $n-1$ components of a vector v in $\mathbb{R}^n$. Then we have

$$\langle \hat{s}, \hat{f}(t, x, y_1^*, z_1)\rangle + s^n f^n(t, x, y^*, z_2) \geqq \langle s, f(t, x, y^*, z^*)\rangle$$
$$\geqq \langle \hat{s}, \hat{f}(t, x, y_1, z_1^*)\rangle + s^n f^n(t, x, y, z_2^*).$$

Taking $z = (z_1^*, z_2)$, in the first inequality, $y = (y_1^*, y_2)$, in the second, and using $s^n > 0$, we obtain

$$f^n(t, x, y^*, z_2) \geqq f^n(t, x, y^*, z_2^*) \geqq f^n(t, x, y_1^*, y_2, z_2^*), \quad \text{for all } z_2 \in Z_2, \text{ and } y_2 \in Y_2.$$

By assumption (b), the right side can be made nonnegative by an appropriate choice of y_2. Therefore, we have $f^n(t, x, y^*, z_2) \geqq 0$ for all $z_2 \in Z_2$, as desired. □

Let $(t_0, x_0) \in [0, T] \times E$, and $v_0 = w(t_0, x_0)$. Then $v_0 \in \mathbb{R}$ by Remark 2 above. Let $U_0(t, x)$ be the extremal pointing strategy defined with respect to $C^+(v_0)$. Note that U_0 can be used to define an extremal strategy Γ_0 in a game with initial point (t_1, x_1) as long as $t_1 \geqq t_0$. Since $C^+(v_0) \subset [0, T] \times E$, it follows from Lemma 1.5, that, for any motion $\varphi \in \Phi[\cdot, t_1, x_1, \Gamma_0]$, $\varphi[t] \in E$ for all $t \in [t_1, T]$. If **H4** holds, then more is true, namely the following lemma.

LEMMA 3.2. *Let (t_0, x_0) and v_0 be as above. If* **H4** *holds and $(t_1, x_1) \in [t_0, T] \times E$, not necessarily in $C^+(v_0)$, then for any motion $\varphi \in \Phi[\cdot, t_1, x_1, \Gamma_0]$, we have $\varphi[t] \in E$ for $t \in [t_1, T]$.*

Proof. Let $\varphi \in \Phi[\cdot, t_1, x_1, \Gamma_e]$ with $\{\varphi_m\}$ as the mth-stage trajectories converging to φ. Suppose there exists a $\acute{t} \in [t_1, T]$ such that $\varphi^n[\acute{t}] < 0$. Then, as in Proposition 3.1, letting $\bar{t} = \inf\{t < \acute{t}: \varphi^n[\tau] < 0 \text{ for all } \tau \in (t, \acute{t}]\}$ and $\sigma = \acute{t} - \bar{t}$ we have

$$\varphi^n[\bar{t}] = 0, \qquad \varphi^n[\tau] < 0 \quad \forall \tau \in [\bar{t}, \bar{t} + \sigma]. \tag{3.50}$$

Let $R > 0$ be such that $|\varphi_m(t)| \leqq R$ for all $t \in [t_1, T]$ and all m. For each m, let τ_k and τ_j be defined as in Proposition 3.1, with $\bar{t} + \sigma$ in place of $\bar{t} + \delta$. Then, just as in that proof,

$$\tau_k \to \bar{t}, \ \tau_j \to \bar{t} + \sigma \quad \text{as } m \to \infty,$$
$$\varphi_m^n(\tau_k) \to 0 \quad \text{as } m \to \infty,$$
$$\varphi_m^n(t) < 0 \quad \forall t \in [\tau_k, \tau_j], \quad \text{for } m \text{ sufficiently large.}$$

Let $\omega(r)$ and the constant M be defined as in Proposition 3.1. Now, for $t \in [\tau_i, \tau_{i+1})$, $k \leqq i < j$,

$$f^n(t, \varphi_m(t), y_i, v_m(t)) \geqq f^n(\tau_i, \varphi_m(\tau_i), y_i, v_m(t)) - \omega((M+1)\delta_m),$$

where $\delta_m = (T - t_0)/m$ and $y_i = U^\alpha(\tau_i, \varphi_m(\tau_i))$, $\alpha = w(t_1, x_1)$. Hence, by **H4**-(ii)

$$f^n(\tau_i, \varphi_m(\tau_i), y_i, z) \geqq 0 \quad \forall z \in Z.$$

Therefore,

$$\varphi_m^n(t) \geqq \varphi_m^n(\tau_i) - \delta_m \omega((M+1)\delta_m), \quad \forall t \in [\tau_i, \tau_{i+1}).$$

Using this, we obtain for $t \in [\tau_k, \tau_j]$,

$$\varphi_m^n(t) \geqq \varphi_m^n(\tau_k) - (j-k)\delta_m \omega((M+1)\delta_m)$$
$$\geqq \varphi_m^n(\tau_k) - m\delta_m \omega((M+1)\delta_m).$$

Letting $m \to \infty$, we get

$$\varphi^m[t] \geqq 0 \quad \forall t \in [\bar{t}, \bar{t} + \sigma],$$

contradicting (3.50). □

THEOREM 3.2. *Let* **H1** *and* **H4** *hold. Then* $w(\cdot,\cdot)$ *is continuous on* $[0, T]\times E$ *if* $g(\cdot)$ *is continuous on* E, *and* $w(\cdot,\cdot)$ *is locally Lipschitz continuous on* $[0, T]\times E$ *if* $g(\cdot)$ *is locally Lipschitz on* E.

Proof. We consider the assertion about Lipschitz continuity. The modifications needed for the other statement are straightforward. Let $X \subset E$ be compact. We show that there exists a constant $K>0$ such that for all (t_0, x_0) and (t_1, x_1) in $[0, T]\times X$,

$$|w(t_0, x_0) - w(t_1, x_1)| \leqq K(|t_0 - t_1| + |x_0 - x_1|).$$

Let $v_0 = w(t_0, x_0)$, $v_1 = w(t_1, x_1)$. Suppose $t_1 \geqq t_0$. Let $\bar{\Gamma}_0, \Gamma_1$ be defined on $[t_1, T]$ extremally with respect to $C^+(v_0)$ and $C^+(v_1)$, respectively. Note that because of **H4**, v_0 and v_1 are finite (see Remark 2). Define

$$d_1 \equiv d((t_1, x_1), C^+(v_0) \cap H_{t_1}).$$

Then, by Lemma 2.1 with $\Omega = C^+(v_0)$, there exists a $C>0$ such that for any $\varphi \in \Phi[\cdot, t_1, x_1, \bar{\Gamma}_0]$,

$$d((t, \varphi[t]), C^+(v_0) \cap H_t) \leqq Cd_1 \quad \forall t \in [t_1, T]. \tag{3.51}$$

Specializing to $t = T$, we have that for any $\varphi \in \Phi[\cdot, t_1, x_1, \bar{\Gamma}_0]$ there exists x_φ such that $(T, x_\varphi) \in C^+(v_0)$ and $|x_\varphi - \varphi[T]| \leqq Cd_1$. It is easily verified (see end of proof) that there exists a constant $\mu>0$, independent of (t_i, x_i) $(i=0, 1)$, for which we have

$$d_1 \leqq \mu(|t_0 - t_1| + |x_0 - x_1|). \tag{3.52}$$

Hence

$$|x_\varphi - \varphi[T]| \leqq C\mu(|t_0 - t_1| + |x_0 - x_1|).$$

It follows that

$$|g(x_\varphi) - g(\varphi[T])| \leqq C_g C\mu(|t_0 - t_1| + |x_0 - x_1|),$$

where C_g is the Lipschitz constant of g on a sufficiently large compact set. Since $g(x_\varphi) \geqq v_0$, we have

$$g(\varphi[T]) \geqq v_0 - C_g C\mu(|t_0 - t_1| + |x_0 - x_1|).$$

By Lemma 3.2, $\varphi[t] \in E$ for all $t \in [t_1, T]$. Therefore, $V(\varphi) = g(\varphi[T])$. Since $\varphi \in \Phi[\cdot, t_1, x_1, \bar{\Gamma}_0]$ was arbitrary, we obtain

$$\inf_{\Delta} P[t_1, x_1, \bar{\Gamma}_0, \Delta] \geqq v_0 - C_g C\mu(|t_0 - t_1| + |x_0 - x_1|).$$

Hence, taking the sup over all Γ's on the left, we have

$$w(t_1, x_1) \geqq v_0 - C_g C\mu(|t_0 - t_1| + |x_0 - x_1|). \tag{3.53}$$

To obtain a similar relation with (t_0, x_0) and (t_1, x_1) interchanged, we proceed as follows. Note that we may not be able to define a strategy over $[t_0, T]$ extremally with respect to $C^+(v_1)$ since $t_1 \geqq t_0$. Instead, let Γ_0 be defined on $[t_0, T]$ extremally with respect to $C^+(v_0)$. Define $\bar{\Gamma}_1$ as the concatenation of Γ_0 on $[t_0, t_1]$ and Γ_1 on $[t_1, T]$. Then given a motion $\bar{\varphi} \in \Phi[\cdot, t_0, x_0, \bar{\Gamma}_1]$; there exists $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma_0]$ such that

$$\text{(a)} \quad \bar{\varphi} \restriction [t_0, t_1] \equiv \varphi \restriction [t_0, t_1] \tag{3.54}$$

$$\text{(b)} \quad \bar{\varphi} \restriction [t_1, T] \in \Phi[\cdot, t_1, \bar{\varphi}[t_1], \Gamma_1].$$

Applying Lemma 3.2 to φ, we have that $\varphi[t] \in E$ for $t \in [t_0, t_1]$. In particular, by (3.54(a)), $\bar{\varphi}[t_1] \in E$. Also, again using Lemma 3.2 and (3.54(b)), $\bar{\varphi}[t] \in E$ for all

$t \in [t_1, T]$. Thus, we have $\bar{\varphi}[t] \in E$, for all $t \in [t_0, T]$. Now define $d_2 = d((t_1, \bar{\varphi}[t_1]), C^+(v_1) \cap H_{t_1})$. It can be shown (see end of proof) that regardless of the choice of $\bar{\varphi}$,

$$d_2 \leqq \mu(|t_0 - t_1| + |x_0 - x_1|). \tag{3.55}$$

Using (3.54(b)), it follows from Lemma 2.1, with $\Omega = C^+(v_1)$, that there exists $x_{\bar{\varphi}}$ such that $(T, x_{\bar{\varphi}}) \in C^+(v_1)$ and

$$|\bar{\varphi}[T] - x_{\bar{\varphi}}| \leqq C d_2 \leqq C\mu(|t_0 - t_1| + |x_0 - x_1|).$$

Hence

$$|g(\bar{\varphi}[T]) - g(x_{\bar{\varphi}})| \leqq C_g C\mu(|t_0 - t_1| + |x_0 - x_1|).$$

Since $g(x_{\bar{\varphi}}) \geqq v_1$,

$$g(\bar{\varphi}[T]) \geqq v_1 - C_g C\mu(|t_0 - t_1| + |x_0 - x_1|).$$

Since the choice of $\bar{\varphi}$ was arbitrary, we get, as before,

$$w(t_0, x_0) \geqq v_1 - C_g C\mu(|t_0 - t_1| + |x_0 - x_1|). \tag{3.56}$$

Let $K = C_g C\mu$. Then, from (3.53) and (3.56), we have the desired conclusion. It remains to verify (3.52) and (3.55).

Proof of (3.52). Take $\varphi \in \Phi[\cdot, t_0, x_0, \Gamma_0]$ and let $\{\varphi_m(\cdot)\}$ be the sequence of mth-stage trajectories converging to φ. Since by Lemma 1.5, $(t_1, \varphi[t_1]) \in C^+(v_0)$, we have

$$\begin{aligned} d_1 &\leqq |(t_1, x_1) - (t_1, \varphi[t_1])| = \lim_{m \to \infty} |(t_1, x_1) - (t_1, \varphi_m(t_1))| \\ &\leqq \lim_{m \to \infty} \{|(t_1, x_1) - (t_0, x_{0_m})| + |(t_0, x_{0_m}) - (t_1, \varphi_m(t_1))|\}, \end{aligned} \tag{3.57}$$

where $x_{0_m} = \varphi_m(t_0)$. Then, for any m,

$$|\varphi_m(t_1) - x_{0_m}| \leqq \int_{t_0}^{t_1} |f(t, \varphi_m(t), u_m(t), v_m(t))|\, dt \leqq M|t_1 - t_0|,$$

where $M = \max\{|f(t, x, y, z)|: t \in [0, T], |x| \leqq R, y \in Y, z \in Z\}$, and $R > 0$ is sufficiently large so that any trajectory with initial point in $[0, T] \times X$ lies in $B_R(0)$. Hence, together with (3.57), we have

$$d_1 \leqq |(t_0, x_0) - (t_1, x_1)| + \sqrt{M^2 + 1}|t_0 - t_1| \leqq \mu(|t_0 - t_1| + |x_0 + x_1|),$$

where $\mu = 1 + \sqrt{M^2 + 1}$. Note that μ depends on X (through M) but not on the initial points (t_0, x_0) and (t_1, x_1).

Proof of (3.56). Since $(t_1, x_1) \in C^+(v_1)$,

$$d_2 \leqq |(t_1, \bar{\varphi}[t_1]) - (t_1, x_1)| \leqq \lim_{m \to \infty} |(t_1, \bar{\varphi}_m(t_1)) - (t_1, x_1)|,$$

where $\{\bar{\varphi}_m\}$ is the sequence of mth-stage trajectories converging to $\bar{\varphi}$. As in the proof of (3.52), above, if $x_{0_m} = \bar{\varphi}_m(t_0)$,

$$\begin{aligned} d_2 &\leqq \lim_{m \to \infty} \{|(t_1, \bar{\varphi}_m(t_1)) - (t_0, x_{0_m})| + |(t_0, x_{0_m}) - (t_1, x_1)|\} \\ &\leqq \sqrt{M^2 + 1}|t_1 - t_0| + |(t_0, x_0) - (t_1, x_1)| \\ &\leqq \mu(|t_0 - t_1| + |x_0 - x_1|). \end{aligned}$$

This completes the proof of Theorem 3.2. □

Remark 4. (Restrictions on the minimizer). If the phase restriction $x^n \geqq 0$ is placed on the minimizer rather than the maximizer, then we clearly need to adjust conditions **H2**, **H3**, and **H4** accordingly. We record these changes here for later reference:
H2 (for minimizer): for all $t \in [0, T]$, $x \in \partial E$

$$\max_z \min_y f^n(t, x, y, z) \geqq 0.$$

H3 (for minimizer) is the same as **H2** with strict inequality.
H4 (for minimizer):

(i) There exists an $x \in E$ such that $w(0, x) > -\infty$.

(ii) If $t \in [0, T)$, $x \notin E$ and $\alpha \in \mathbb{R}$ are such that $V^\alpha(t, x)$, the extremal strategy with respect to $C^-(\alpha)$, is defined then

$$\min_y f^n(t, x, y, z^*(t, x)) \geqq 0,$$

where $z^*(t, x)$ is the outcome of $V^\alpha(t, x)$.

Remark 5 (More general sets E). The arguments used to prove Proposition 3.1, and Theorems 3.1 and 3.2 are essentially of a local nature. Using this fact, we can extend these theorems to the setting where E is a closed set with nonempty interior and C^2 boundary. In doing so, it is only necessary to extend Lemmas 3.1 and 3.2 to the new setting. Now, through a change of coordinates, E can be locally mapped onto a subset of $\{x: x^n \geqq 0\}$. We then apply the results obtained above the problem in the new coordinates.

Clearly assumptions **H2**, **H3**, and **H4** have to be expressed in terms different from before. Such adjustments are straightforward; for example, **H3** would read:
H3': If $t \in [0, T]$, $x \in \partial E$ and $n(x)$ is the inward normal to ∂E at x, then

$$\max_y \min_z \langle n(x), f(t, x, y, z)\rangle > 0.$$

In a similar manner, these theorems extend to sets E of form $E = E_1 \cap E_2 \cap \cdots \cap E_k$, where each of the E_i's is a closed set with $\operatorname{int}(E_i) \neq \varnothing$ and C^2 boundary with the further restriction that if $x \in E_1 \cap E_2 \cap \cdots \cap E_k$, and $n_i(x) =$ the (inward) normal to ∂E_i at x, then the set $\{n_i(x): 1 \leqq i \leqq k\}$ is a linearly independent set of vectors. This covers, for example, the case of $E = \{x \in \mathbb{R}^n: x^1 \geqq 0, \cdots, x^k \geqq 0\}$. We refer the interested reader to § 4 of [6] for the details.

Our last result differs from Theorems 3.1 and 3.2 in that we make no assumptions on the shape or smoothness of the phase set E.

THEOREM 3.3. *Suppose w is continuous on $[0, T] \times \partial E$, then w is continuous on $[0, T] \times E$.*

Before starting the proof, we make an observation. Suppose $(t_0, x_0) \in C^+(\alpha)$, for some $\alpha \in \mathbb{R}$, and $t_1 \geqq t_0$. Referring to the proof of Lemma 2.1, we note that given a compact set $X \subset E$ containing x_0, if Γ_e is the strategy defined on $[t_1, T]$, extremally with respect to $C^+(\alpha)$, then there exist constants β, K and a function $E(\cdot)$, all depending only on f and X, such that for any $\varphi_m \in \Phi_m(\cdot, t_1, x_1, \Gamma_e)$, $x_1 \in X$, we have, for all $t \in [t_1, T]$,

$$d^2((t, \varphi[t]), C^+(\alpha) \cap H_t) \leqq e^{\beta K} d^2((t_1, x_1), C^+(\alpha) \cap H_{t_1}) + E(\delta_m)(e^{\beta K - 1}/\beta),$$

where $\delta_m = (T - t_1)/m$. It follows that if $C > e^{(\beta K)/2}$ then there exists an m^*, independent of the set $C^+(\alpha)$, such that for any $\varphi_m \in \Phi_m(\cdot, t_1, x_1, \Gamma_e)$, if $m \geqq m^*$, then

$$d((t, \varphi_m(t)), C^+(\alpha) \cap H_t) \leqq C d((t_1, x_1), C^+(\alpha) \cap H_{t_1}), \quad \forall t \in [t_1, T].$$

Proof of the Theorem. Let $(t_0, x_0) \in [0, T] \times E$. By Lemma 1.2, we only need to show that w is lower semicontinuous; i.e., $\forall \varepsilon > 0$, $\exists \delta > 0$ such that

$$|t_0 - t_1| + |x_0 - x_1| < \delta \Rightarrow w(t_1, x_1) > w(t_0, x_0) - \varepsilon. \tag{3.58}$$

Let G be a bounded neighborhood of x_0 and $R > 0$ be such that any trajectory with initial point in $[0, T] \times \bar{G}$ is contained in $B_R(0)$. Since w is, by hypothesis, continuous on $[0, T] \times \partial E$, there exists a $\sigma > 0$ such that for any t, $t' \in [0, T]$ and x, $x' \in \partial E \cap B_R(0)$,

$$|t - t'| + |x - x'| < \sigma \Rightarrow \| w(t, x) - w(t', x')| < \varepsilon/2. \tag{3.59}$$

Case 1. $t_1 \geqq t_0$. Define, for any $x_1 \in \bar{G}$, a strategy $\Gamma(t_1, x_1)$ as follows. The mth-partition is a uniform subdivision $\{t_1 < \tau_1 < \cdots < \tau_m = T\}$ of $[t_1, T]$. The mth-stage strategy is defined positionally as follows. Let $\Gamma_{m,0} = U_0(t_1, x_1)$ where $U_0 = U_e(C^+(v_0))$, and $v_0 = w(t_0, x_0)$. Suppose $u_m(\cdot)$, the outcome of Γ_m versus some Δ_m, has been defined on $[t_1, \tau_i)$ and φ_m is the corresponding trajectory on $[t_1, \tau_i]$. Let $v_m(\cdot)$ be the outcome of Δ_m on $[t_1, \tau_i)$. Then if $\varphi_m(\tau_i) \in E$ define $\Gamma_{m,i+1}(u_m, v_m) = U_0(\tau_i, \varphi_m(\tau_i))$. Suppose $\kappa = \kappa(m)$ is the smallest index such that $\varphi_m(\tau_k) \notin E$. Define $t_m = \inf\{t' < \tau_k : \varphi_m(\tau) \notin E$, for all $\tau \in (t', \tau_\kappa]\}$. Note that $t_m \in [\tau_{\kappa-1}, \tau_\kappa)$, by its definition, and $\varphi_m(t_m) \in \partial E$. Let $v_m = w(t_m, \varphi_m(t_m))$ and $U_m = U_e(C^+(v_m))$. Now, define $\Gamma_{m,i+1}(u_m, v_m) = U_m(\tau_i, \varphi_m(\tau_i))$, for all $i \geqq \kappa$.

It follows from the above observation, with $X = \bar{G}$, that if φ_m is an mth-stage trajectory of $\Gamma(t_1, x_1)$, then

$$d((t, \varphi_m(t)), C^+(v_0) \cap H_t) \leqq C d((t_1, x_1), C^+(v_0) \cap H_{t_1}), \quad \forall t \in [t_1, \tau_\kappa], \tag{3.60}$$

and

$$d((t, \varphi_m(t)), C^+(v_m) \cap H_t) \leqq C d((\tau_\kappa, \varphi_m(\tau_k)), C^+(v_m) \cap H_{\tau_\kappa}), \quad \forall t \in [\tau_\kappa, T]. \tag{3.61}$$

Now recall that, as observed in the proof of Theorem 3.2 (cf. (3.52)), there exists a constant μ, depending on X, such that

$$d((t_1, x_1), C^+(v_0) \cap H_{t_1}) \leqq \mu\{|t_0 - t_1| + |x_0 + x_1|\}, \tag{3.62}$$

$$d((\tau_\kappa, \varphi_m(\tau_\kappa)), C^+(v_m) \cap H_{\tau_\kappa}) \leqq \mu\{|\tau_\kappa - t_m| + |\varphi_m(t_m) - \varphi_m(\tau_\kappa)|\}. \tag{3.63}$$

From (3.62) and (3.63), we have that for all $m \geqq m^*$,

$$d((t, \varphi_m(t)), C^+(v_0) \cap H_t) \leqq C\{|t_0 - t_1| + |x_0 - x_1|\}, \quad \forall t \in [t_1, \tau_\kappa]. \tag{3.64}$$

Let $\mathcal{F} = C^+(v_0) \cap [t_0, T] \times (\partial E \cap \bar{B}_R(0))$. $\mathcal{F}$ is compact, being bounded and closed. For each m, define $\rho_m \in \mathbb{R}$ as follows. If $\mathcal{F} = \varnothing$ or if $\varphi_m(t) \in E$, for all $t \in [t_1, T]$ and all $\varphi_m(t) \in \Phi_m(\cdot, t_1, x_1, \Gamma(t_1, x_1))$ then $\rho_m = 0$. Otherwise $\rho_m = \sup\{d((t_m, \varphi_m(t_m)), \mathcal{F})$: $\varphi_m \in \Phi_m(\cdot, t_1, x_1, \Gamma(t_1, x_1))\}$, where t_m is as defined above.

Claim. For every positive number σ, there exists a $\delta > 0$ such that if $|t_0 - t_1| + |x_0 - x_1| \leqq \delta$, $t_1 \geqq t_0$, then $\rho_m \leqq \sigma$ for all $m \geqq m^*$. To prove this statement suppose, on the contrary, that there exist $\sigma > 0$, sequences $(t_n, x_n) \to (t_0, x_0)$, $t_n \geqq t_0$, and $\{m(n)\}$ such that $\rho_{m(n)} > \sigma$ for every n. By the definition of ρ_m, we obtain a sequence $\{\varphi_{m(n)}\}$ with

$$d((t_{m(n)}, \varphi_{m(n)}(t_{m(n)})), \mathcal{F}) \geqq \sigma, \quad \forall n. \tag{3.65}$$

Applying (3.64) with $t = t_{m(n)}$, we have that $\exists z_n$ such that $(t_{m(n)}, z_n) \in C^+(v_0)$ and

$$|z_n - \varphi_{m(n)}(t_{m(n)})| \leqq C(|t_n - t_0| + |x_0 - x_n|). \tag{3.66}$$

It follows that $\{z_n\}$ is bounded. Hence, by taking a subsequence, we may assume that there exist t^* and z^* such that $t_{m(n)} \to t^*$ and $z_n \to z^*$. Therefore, from (3.66) we get that $\varphi_{m(n)}(t_{m(n)}) \to z^*$, as $n \to \infty$. Since $\varphi_{m(n)}(t_{m(n)}) \in \partial E$ for all n, we conclude that $z^* \in \partial E$. Also, since $C^+(v_0)$ is closed, we have $(t^*, z^*) \in C^+(v_0)$. Therefore $(t^*, z^*) \in \mathcal{F}$. But now $\varphi_{m(n)}(t_{m(n)}) \to z^*$ and $t_{m(n)} \to t^*$ contradict (3.65).

We now prove (3.58) for this case. Let σ be as chosen in (3.59). By the above claim there exists $\delta_1 > 0$ such that if $|t_0 - t_1| + |x_0 - x_1| \leqq \delta_1$ then $\rho_m \leqq \sigma$. Now let φ be any motion in $\Phi[\cdot, t_1, x_1, \Gamma(t_1, x_1)]$ with $\{\varphi_m\}$ as its mth-stage trajectories. Without loss of generality, we may assume that $\varphi_m(t_1) = x_1$ (i.e., $\varphi_m \in \Phi_m(\cdot, t_1, x_1, \Gamma(t_1, x_1))$). We claim that $\varphi[t] \in E$, for all $t \in [t_1, T]$ and $g(\varphi[T]) \geqq v_0 - \varepsilon$.

Verification. Suppose that for all m sufficiently large, $\kappa = \kappa(m) < m$ (κ as defined above); otherwise the desired conclusion follows from (3.64) and the continuity of g. Recall that $t_m \in [\tau_{\kappa-1}, \tau_\kappa)$. Hence, since the norm of the partition tends to zero as $m \to \infty$, we get

$$|\tau_\kappa - t_m| + |\varphi_m(t_m) - \varphi_m(\tau_k)| \to 0 \quad \text{as } m \to \infty. \tag{3.67}$$

It follows from (3.61), (3.63), (3.67) and the fact that, by the definition of v_m, $(t_m, \varphi_m(t_m)) \in C^+(v_m)$, that

$$\lim_{m\to\infty} \max_{t\in[\tau_\kappa, T]} d((t, \varphi_m(t)), C^+(v_m) \cap H_t) = 0. \tag{3.68}$$

Hence, by the continuity of $g(\cdot)$, for all m sufficiently large,

$$g(\varphi_m(T) \geqq v_m - \varepsilon/4. \tag{3.69}$$

Note also that by the definition of τ_κ, and since $t_m \in [\tau_{\kappa-1}, \tau_\kappa)$,

$$\max_{t\in[t_1, \tau_\kappa]} d(\varphi_m(\tau), E) = \max_{t\in[t_m, \tau_\kappa]} d(\varphi_m(\tau), E) \to 0, \quad \text{as } m \to \infty.$$

Furthermore, since $C^+(v_m) \subset [0, T] \times E$, we get from (3.68), that

$$\max_{t\in[\tau_\kappa, T]} d(\varphi_m(t), E) \to 0, \quad \text{as } m \to \infty.$$

Hence, $\varphi[t] \in E$, for all $t \in [t_1, T]$. Therefore, $V(\varphi) = g(\varphi[T])$, and using (3.69),

$$V(\varphi) = g(\varphi[T]) \geqq v_m - \varepsilon/2, \tag{3.70}$$

for all m sufficiently large. Now, by the choice of δ_1, for every m, there exists $(\bar{t}_m, \bar{x}_m) \in \mathscr{F}$ such that $|\bar{t}_m - t_m| + |\varphi_m(t_m) - \bar{x}_m| \leqq \sigma$. Therefore, by (3.59) and the fact that $\mathscr{F} \subset C^+(v_0)$, we have

$$v_m = w(t_m, \varphi_m(t_m)) \geqq w(\bar{t}_m, \bar{x}_m) - \varepsilon/2 \geqq v_0 - \varepsilon/2. \tag{3.71}$$

Thus, by (3.70), $V(\varphi) \geqq v_0 - \varepsilon$. Since $\varphi \in \Phi[\cdot, t_1, x_1, \Gamma(t_1, x_1)]$ was arbitrary, taking inf over all φ's and sup over all Γ's, we have $w(t_1, x_1) \geqq w(t_0, x_0) - \varepsilon$, as desired. This completes Case 1.

Let us note that we have actually shown, above, that for an initial point (t_1, x_1), with $t_1 \geqq t_0$, if the first player plays that mth-stage of $\Gamma(t_1, x_1)$ at the mth-stage, then any for $\varphi_m \in \Phi_m(\cdot, t_1, x_1, \Gamma(t_1, x_1))$, we have

$$g(\varphi_m(T)) \geqq v_0 - 3\varepsilon/4, \tag{3.72}$$

(cf., (3.69) and (3.71)) for all m sufficiently large, provided $|t_0 - t_1| + |x_0 - x_1| < \delta_1$.

Case 2. $t_1 < t_0$. Let $M = \max\{1, |f(t, x, y, z)|: t \in [0, T], |x| \leqq R, y \in Y, z \in Z\}$. Let $\delta_2 = \min\{\delta_1, \sigma\}/(M+1)$. Suppose that $|t_0 - t_1||x_0 - x_1| < \delta_2$. Define a strategy $\Gamma(t_1, x_1)$ as follows. The mth-stage partition is $\Pi = \{t_1 < \tau_1 < \cdots < \tau_m = t_0 < \cdots < \tau_{2m} = T\}$ where $|\tau_i - \tau_{i-1}| = (t_0 - t_1)/m$ if $i \leqq m$ and $|\tau_i - \tau_{i-1}| = (T - t_0)/m$ if $i > m$. For the mth-stage strategy, take $\Gamma_{m,0} = U_1(t_1, x_1)$, where $U_1 = U_e(C^+(w(t_1, x_1)))$. Suppose $u_m(\cdot)$ is the outcome of $\Gamma_m(t_1, x_1)$ versus some Δ_m defined on $[t_1, \tau_i)$ and $\varphi_m(\cdot) = \varphi_m(\cdot, t_1, x_1, u_m, v_m)$ defined on $[t_1, \tau_i]$. If $\varphi_m(\tau_i) \in E$, then

$$\Gamma_{m,i+1}(u_m, v_m) = \begin{cases} U_1(\tau_i, \varphi_m(\tau_i)) & \text{if } i < m \\ U_0(\tau_i, \varphi_m(\tau_i)) & \text{if } i \geqq m. \end{cases}$$

Let $\kappa = \kappa(m)$ be the smallest index such that $\varphi_m(\tau_\kappa) \notin E$. Then define $t_m \in [\tau_{\kappa-1}, \tau_\kappa)$ as before, let $v_m = w(t_m, \varphi_m(t_m))$ and $U_m = U_e(C^+(v_m))$. Now, define $\Gamma_{m,i+1}(u_m, v_m) = U_m(\tau_i, \varphi_m(\tau_i))$, for all $i \geqq \kappa$. Note that if $\varphi_m \in \Phi_m(\cdot, t_1, x_1, \Gamma(t_1, x_1))$, then

$$
\begin{aligned}
|t - t_0| + |\varphi_m(t) - x_0| &\leqq |t - t_0| + |\varphi_m(t) - \varphi_m(t_1)| + |\varphi_m(t_1) - x_0| \\
&\leqq (M+1)(|t_0 - t_1| + |x_0 - x_1|) < \delta_2, \quad \text{for all } t \in [t_1, t_0].
\end{aligned} \tag{3.73}
$$

Now if $\varphi_m(t) \in E$ for all $t \in [t_1, t_0]$, then, because of the definition of $\Gamma(t_1, x_1)$, we have that (3.72) holds. On the other hand, suppose that for some $t \in [t_1, t_0]$, $\varphi_m(t) \notin E$, then $k \leqq m$. We may assume, in this case, that $x_0 \in \partial E$, for if $x_0 \in \operatorname{int}(E)$ then, by shrinking δ_2, if necessary, we can arrange to have $\varphi_m(t) \in \operatorname{int}(E)$ for all $t \in [t_1, t_0]$. Now, by arguments similar to those of Case 1, we have $g(\varphi_m(T) \geqq v_m - 3\varepsilon/4$, (cf., (3.69)), for all m sufficiently large. But, by (3.73), $|t_m - t_0| + |\varphi_m(t) - x_0| < \delta_2 < \sigma$. Hence, by (3.59), we have

$$
v_m = w(t_m, \varphi_m(t_m)) \geqq w(t_0, x_0) - \varepsilon/2.
$$

Thus, (3.72) holds in this case also. Note that, just as in Case 1, using the fact that sets $C^+(\alpha) \subset [0, T] \times E$ for any $\alpha \in \mathbb{R}$, we have that if $\varphi \in \Phi[\cdot, t_1, x_1, \Gamma(t_1, x_1))$ then $\varphi[t] \in E$ for all $t \in [t_1, T]$. Hence it follows from (3.72) and the continuity of $g(\cdot)$ that

$$
V(\varphi) = g(\varphi[T]) \geqq v_0 - \varepsilon.
$$

Taking inf over all φ's and sup over all Γ's we obtain $w(t_1, x_1) \geqq w(t_0, x_0) - \varepsilon$ in this case. Therefore if $\delta = \min(\delta_1, \delta_2)$, we have (3.58) as desired. This completes the proof of the Theorem. □

Example (Battle of Bunker Hill). The dynamics are:

$$
\begin{aligned}
\dot{x}^1 &= -c_1 p_1 x^3 z \\
\dot{x}^2 &= -c_1 z \\
\dot{x}^3 &= -c_2 p_2 x^1 y \\
\dot{x}^4 &= -c_2 y
\end{aligned}
$$

where the c_i's and p_i's are positive constants. The control sets are $Y = Z = [0, 1]$, and $T =$ some given final time. The terminal payoff is defined through either of the following functions: $g_1(x) = x^1 - x^3$ and $g_2(x) = x^1 x^4 - x^3 x^2$. This game was studied heuristically by Isaacs in [7] (cf. [7] for an interpretation of the game). One of the players, say player I, is restricted to $E = \{x: x^3 \geqq 0, x^4 \geqq 0\}$. It is assumed further that the initial point of the game is such that $x^1 \geqq 0$ and $x^2 \geqq 0$ are automatically satisfied throughout the game. It is then easy to check that if $x_0 \in \partial E$, then the only admissible control for player I is $y \equiv 0$. Hence, for $x_0 \in \partial E$ the game reduces to an optimal control problem for player II. Its value is easily seen to be continuous. Therefore, by Theorem 3.3, the value of the original game, which exists by Theorem 1.1, is also continuous.

Remark 6 (Hamilton-Jacobi-Isaacs' equation). In Theorems 3.1, 3.2, and 3.3, we have given conditions under which $w(\cdot, \cdot)$ will be continuous. By essentially the same arguments as those of [2], but not using a Lipschitz continuity assumption on W, it follows (cf. [6]) from the stability of $C^+(\alpha)$ and Lemma 1.6 that w satisfies the following.

(1) If φ is continuously differentiable on a neighborhood of $[0, T] \times E$ and $w - \varphi$ has a local max at $(t_0, x_0) \in (0, T) \times E$, then

$$
\varphi_t(t_0, x_0) + H(t_0, x_0, D\varphi(t_0, x_0)) \geqq 0;
$$

(2) If φ is continuously differentiable on a neighborhood of $[0, T]\times E$ and $w-\varphi$ has a local min at $(t_0, x_0)\in(0, T)\times \underline{\text{int}(E)}$, then

$$\varphi_t(t_0, x_0)+H(t_0, x_0, D\varphi(t_0, x_0))\leqq 0,$$

where $H(t, x, p)=\max_y \min_z \langle p, f(t, x, y, z)\rangle$ and $D\varphi=(\partial\varphi/\partial x^1, \cdots, \partial\varphi/\partial x^n)$.

Let us note that conditions (1) and (2) are precisely the ones proposed by Soner in [12] as defining "constrained viscosity solutions" of the Hamilton-Jacobi equation. Under a mild restriction on ∂E, by combining the methods of Ishii ([13]) and Soner ([12]), we can show (cf., [6]) that (1), (2), and the terminal condition $w(T, x)=g(x)$, for all $x\in E$, determine w uniquely on $[0, T]\times E$. We note that this type of uniqueness theorem is not new. The first such result was obtained by Soner (cf., [12]). More recently, Capuzzo-Dolcetta and Lions have given a detailed treatment of viscosity solutions for problems with state contraints in [15] (see also [14]). In this paper they obtain, in addition to many other interesting results, uniqueness theorems of the type mentioned above for more general functions H (cf., [15, Thms. III.1-III.4]).

Acknowledgment. This work comprises parts of the author's Ph.D. thesis. The author would like to thank his major advisor, Professor L. D. Berkovitz, for suggesting the problem and for his subsequent, invaluable suggestions and guidance.

REFERENCES

[1] L. D. BERKOVITZ, *The existence of value and saddle point in games of fixed duration*, SIAM J. Control Optim., 23 (1985), pp. 172-196.

[2] ———, *Characterization of the value of differential games*, Appl. Math. Optim., 17 (1988), pp. 177-183.

[3] ———, *The existence of value and saddle point in games of fixed duration: erratum and addendum*, SIAM J. Control Optim., 26 (1988), pp. 740-742.

[4] A. FRIEDMAN, *Differential Games*, Wiley-Interscience, New York, 1971.

[5] ———, *Differential Games*, CBMS Regional Conference Series in Mathematics 18, American Mathematical Society, Providence, RI, 1974.

[6] K. HAJI-GHASSEMI, Ph.D. Thesis, Purdue University, W. Lafayette, IN, 1988.

[7] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[8] R. C. SCALZO, *Differential games with restricted phase coordinates*, SIAM J. Control, 12 (1974), pp. 426-434.

[9] A. I. SUBBOTIN, *Differential games with constraints on phase states*, Soviet Math. Dokl. 11 (1970), pp. 933-936.

[10] L. S. ZAREMBA, *Existence of value in differential games with fixed time duration*, J. Optim. Theory Appl., 38 (1982), pp. 581-598.

[11] ———, *Existence of value in differential games with terminal cost function*, J. Optim. Theory Appl., 39 (1983), pp. 89-104.

[12] H. M. SONER, *Optimal control with state space constraint I*, SIAM J. Control Optim., 24 (1986), pp. 552-561.

[13] H. ISHII, *Uniqueness of unbounded viscosity solutions of the Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 721-748.

[14] I. CAPUZZO-DOLCETTA, *Hamilton-Jacobi equations with constraints*, in Stochastic Differential Systems, Stochastic Control Theory and Application, W. Fleming and P. L. Lions, eds., Springer-Verlag, Berlin, New York, 1988, pp. 99-106.

[15] I. CAPUZZO-DOLCETTA AND P. L. LIONS, *Hamilton-Jacobi equations and state-constrained problems*, IMA Preprint Series no. 342, September 1987.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 653-670, May 1990

010

A SKEW TOEPLITZ APPROACH TO THE H^∞ OPTIMAL CONTROL OF MULTIVARIABLE DISTRIBUTED SYSTEMS*

HITAY ÖZBAY† AND ALLEN TANNENBAUM‡

Abstract. In this paper the problem of the H^∞ optimization of multivariable distributed systems in the four block setting is studied. This work is based on several previous papers and employs the skew Toeplitz framework developed in [*Operator Theory: Adv. Appl.*, 32 (1988), pp. 21-43], [*Operator Theory: Adv. Appl.*, 32 (1988), pp. 93-112], [*Operator Theory and Integral Equations*, 11 (1988), pp. 726-767], [*J. Functional Anal.*, 74 (1987), pp. 146-159], [SIAM J. Math. Anal., 19 (1988), pp. 1081-1091].

Key words. H^∞-optimal control, distributed system, four block problem, skew Toeplitz operator

AMS(MOS) subject classifications. 93B35, 93C05

1. Introduction. In the past few years, there has been a major research effort devoted to the study of the H^∞ optimization of linear systems. We refer the reader to [13] for an extensive set of references. In this paper we consider the problem of the H^∞-optimization for multivariable distributed systems.

Motivations leading to the H^∞ optimization in systems theory lie in the most natural problems of control engineering such as robust stabilization, sensitivity minimization, and model matching. It can be shown that, in the sense of H^∞ optimality, these problems are equivalent, and can be stated (see [13]) as one *standard problem.* Consider the setup shown in Fig. 1. In this configuration w, u, y, and z are vector-valued signals with w the exogenous input representing the disturbances, measurement noises, etc., u the command signal, z the output to be controlled, and y the measured output. G represents a combination of the plant and the weights in the control system. The standard H^∞ problem is to find a stabilizing controller K such that the H^∞ norm of the transfer function from w to z is minimized. For finite-dimensional systems an expression for a suboptimal controller is given in [2] and [4] using a state-space approach.

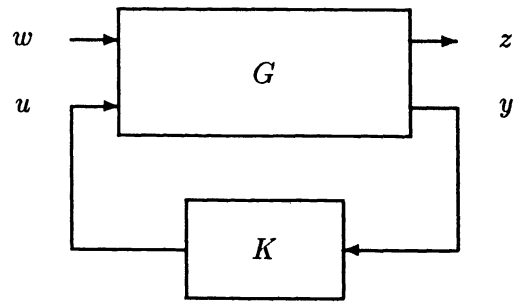


FIG. 1

* Received by the editors November 21, 1988; accepted for publication (in revised form) July 12, 1989. This work was supported by National Science Foundation grants ECS-8704047 and DMS-8811084, and by Air Force Office of Scientific Research grant AFOSR-88-0020.

† Department of Electrical Engineering and Control Sciences and Dynamical Systems Center, University of Minnesota, Minneapolis, Minnesota 55455.

‡ Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455 and Department of Electrical Engineering, Technion, Israel.

Now it is quite well known that an optimal solution of the standard problem can be reduced to finding the singular values of a certain operator (the so-called **four block operator**) that will be defined below. For details we refer the reader to [5]–[7]. Depending on the specific problem considered, the corresponding four block operator can be simplified to a 2-block or a 1-block operator.

This paper is based on several previous papers [6]–[12], [21], and basically employs the skew Toeplitz framework of [3] to study the standard problem. We should note that software for the implementation of the techniques used in this paper has already been written at the Systems Research Center of Honeywell, Minneapolis in collaboration with Blaise Morton, and has been applied to several distributed systems including a flexible beam problem. We plan to write a paper with several such "benchmark" examples with Blaise Morton in the near future.

The present paper is organized as follows. In the next section we set up some notation and give some background on the ideas taken from previous work. In § 3 we derive our main result which is a rank type formula for the singular values of the four block operator. We illustrate a special case of our main result by considering SISO plants in § 4, and by giving an explicit example in § 5. Finally, in § 6 we summarize our results and make some comments.

2. Problem definition and preliminary remarks. We will now state the standard H^∞ problem and define the four block operator. We will also present some preliminary results from earlier work [3], [6], [7]. Throughout the paper all Hardy spaces are defined on the unit disc D in the standard way. For an integer m we denote the canonical unilateral shift (defined by multiplication by z) on $H^2(\mathbf{C}^m)$ by $S: H^2(\mathbf{C}^m) \to H^2(\mathbf{C}^m)$ and the bilateral shift on $L^2(\mathbf{C}^m)$ by $U: L^2(\mathbf{C}^m) \to L^2(\mathbf{C}^m)$. Let W, F, G, J, and M be H^∞ matrices, of sizes $p \times m$, $p \times l$, $q \times m$, $q \times l$, and $p \times p$, respectively, with $p \leqq \max\{m, l\}$, where W, F, G, J have rational entries, and M is a nonconstant inner matrix. These matrices are associated with the weighting matrices and the plant in the usual way of transforming the standard problem to the 4-block framework (i.e., via Youla parametrization and some inner outer factorizations; see, e.g., [13] and [20]). It is important to note that for many problems of interest, in the case of rational weights and distributed stable plants, this reduces to the kind of problem described below. See [15] for all the details. The standard H^∞ problem amounts to finding

$$\mu := \inf\left\{ \left\| \begin{bmatrix} W - MQ & F \\ G & J \end{bmatrix} \right\|_\infty : Q \in H^\infty p \times m \right\},$$

where for a $k \times n$ matrix of the form $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$, ($A$, B, C, D having appropriate sizes with entries in L^∞), we set

$$\left\| \begin{bmatrix} A & B \\ C & D \end{bmatrix} \right\|_\infty = \operatorname{ess\,sup}\left\{ \left\| \begin{bmatrix} A(\zeta) & B(\zeta) \\ C(\zeta) & D(\zeta) \end{bmatrix} \right\| : |\zeta| = 1 \right\}.$$

(For the norm on the right-hand side the $k \times n$ matrix is taken as a linear operator from $\mathbf{C}^n$ to $\mathbf{C}^k$ for each fixed ζ in ∂D, the unit circle.) Note that if $F = G = J = 0$ then this problem reduces to the classical Nehari problem, which is also known as the *1-block problem.* For $F = J = 0$ we have the *2-block problem.*

To the $p \times p$ inner matrix M, we associate the spaces $H(M) := H^2(\mathbf{C}^p) \ominus MH^2(\mathbf{C}^p)$ and $L(M) := L^2(\mathbf{C}^p) \ominus MH^2(\mathbf{C}^p)$. Let $P_{H(M)}: H^2(\mathbf{C}^p) \to H(M)$, $P_{L(M)}: L^2(\mathbf{C}^p) \to L(M)$, $P_{H^2}: L^2(\mathbf{C}^p) \to H^2(\mathbf{C}^p)$, and $P_{L^2 \ominus H^2}: L^2(\mathbf{C}^p) \to L^2(\mathbf{C}^p) \ominus H^2(\mathbf{C}^p)$ be orthogonal projections.

We now define the 4-block operator (see [5] and [7]):

$$A := \begin{bmatrix} P_{H(M)}W(S) & P_{L(M)}F(U) \\ G(S) & J(U) \end{bmatrix}.$$

Note that $A: H^2(\mathbf{C}^m) \oplus L^2(\mathbf{C}^l) \to L(M) \oplus L^2(\mathbf{C}^q)$.

In the paper, by a slight abuse of notation, ζ will denote a complex variable as well as an element of ∂D. The context will make the meaning clear. Note that $W(S)$ can be seen as the operator defined by multiplication by $W(\zeta)$, and similarly for $G(S)$, $F(U)$, and $J(U)$. Using the commutant lifting theorem [18, pp. 257-259], we can show that μ is equal to $\|A\|$. (See [5] and [7] for the details.) Note that $\|A\|^2$ is the largest element of $\sigma(A^*A)$, the spectrum of A^*A. $\sigma(A^*A)$ consists of the discrete spectrum (i.e., eigenvalues with finite multiplicity), which we denote by $\sigma_d(A^*A)$, and its complement $\sigma_e(A^*A)$, the essential spectrum. The essential spectrum of A^*A consists of those $\lambda \in \mathbf{C}$ for which there exists

$$\begin{bmatrix} x_n \\ y_n \end{bmatrix} \in H^2(\mathbf{C}^m) \oplus L^2(\mathbf{C}^l) \quad \text{with} \quad \left\| \begin{bmatrix} x_n \\ y_n \end{bmatrix} \right\|_2 = 1 \quad \forall n \geqq 1,$$

and $\left[\begin{smallmatrix} x_n \\ y_n \end{smallmatrix}\right] \to 0$ weakly as $n \to \infty$, such that

$$(\lambda I - A^*A) \begin{bmatrix} x_n \\ y_n \end{bmatrix} \to 0 \quad \text{as } n \to \infty.$$

The essential norm, denoted by $\|A\|_e$, is defined as

$$\|A\|_e^2 = \max \{\lambda : \lambda \in \sigma_e(A^*A)\}.$$

In the SISO case we have that (see [7, Thm. 3.2])

$$\|A\|_e = \max(\alpha, \beta, \gamma),$$

where

$$\alpha = \max \left\{ \left\| \begin{bmatrix} W(\zeta) & F(\zeta) \\ G(\zeta) & J(\zeta) \end{bmatrix} \right\| : \zeta \in \sigma_e(T) \right\},$$

$$\beta = \max \{ \| [G(\zeta) \quad J(\zeta)] \| : \zeta \in \partial D \},$$

$$\gamma = \max \left\{ \left\| \begin{bmatrix} F(\zeta) \\ J(\zeta) \end{bmatrix} \right\| : \zeta \in \partial D \right\}.$$

$\sigma_e(T)$ denotes the essential spectrum of the operator $T := P_{H(M)}S|_{H(M)}$. We let $\mathcal{R}$ be the set of all $\lambda \in \partial D$ that do not lie on any of the open arcs of ∂D on which $M(\zeta)$ is a unitary operator-valued analytic function. Then from [17] and [18], we have that

$$\sigma_e(T) = \mathcal{R}.$$

In the case of infinite-dimensional MIMO systems it may be difficult to find the essential norm of A. Nevertheless, upper and lower bounds can be obtained in terms of α, β, γ. This is discussed in detail in § 3.2.

Note that when $\|A\| > \|A\|_e$, $\|A\|^2$ is an eigenvalue of A^*A. Here we are going to develop a rank type formula for the eigenvalues of A^*A. We will show that this formula is obtained by a certain linear system of equations (called the *singular system* in [7]). These equations are derived from the inversion of two Toeplitz operators and the essential inversion of a skew Toeplitz operator. It is important to note that in the

2-block problem, one of the Toeplitz operator inversions disappears, and in the 1-block case the same is true for both of the Toeplitz operator inversions. The Fredholm conditions on the invertibility of the skew Toeplitz operator (which is essentially invertible) and the coupling between various systems of equations constitute the singular system. See also [3] and [7].

3. Main results.

3.1. Discrete spectrum. Let us begin with the following assumption $W = B/k$, $F = C/k$, $G = D/k$, and $J = E/k$, where B, C, D, E are polynomial matrices and k is a scalar polynomial. We denote by n an upper bound for the degree of the entries of all polynomial matrices appearing throughout the paper.

Now it is easy to see that ρ^2 is an eigenvalue of A^*A if and only if there exists a nonzero

$$\begin{bmatrix} x \\ y \end{bmatrix} \in H^2(\mathbf{C}^m) \oplus L^2(\mathbf{C}^l),$$

such that

$$\begin{aligned} &(\rho^2 k(S)^* k(S) I - B(S)^* P_{H(M)} B(S) - D(S)^* D(S))x \\ &\quad -(P_{H^2}(B(U)^* P_{L(M)} C(U) + D(U)^* E(U)))y = 0, \end{aligned} \tag{1a}$$

and

$$\begin{aligned} &-((C(U)^* P_{H(M)} B(S) + E(U)^* D(S)))x \\ &\quad +(\rho^2 k(U)^* k(U) I - C(U)^* P_{L(M)} C(U) - E(U)^* E(U))y = 0. \end{aligned} \tag{1b}$$

Note that $P_{H(M)} B(S)x = B(\zeta)x - M(\zeta) P_{H^2} M(\zeta)^* B(\zeta)x$. Following the techniques used in [3], we make the factorization

$$M(\zeta)^* B(\zeta) = \Omega_b(\zeta) M_b(\zeta)^*, \tag{f1}$$

where $\Omega_b(\zeta)$ is a polynomial matrix of size $p \times m$ and $M_b(\zeta)$ is an inner matrix of size $m \times m$. We now decompose the space $H^2(\mathbf{C}^m)$ as $H(M_b) \oplus M_b H^2(\mathbf{C}^m)$, and express $x = x_b + M_b x_b'$ where $x_b \in H(M_b)$ and $x_b' \in H^2(\mathbf{C}^m)$. Then we have

$$P_{H^2} M(\zeta)^* B(\zeta)x = P_{H^2} \Omega_b(\zeta) M_b(\zeta)^* (x_b + M_b x_b').$$

Since M_b is inner,

$$P_{H^2} M(\zeta)^* B(\zeta)x = \Omega_b(\zeta) x_b' + P_{H^2} \Omega_b(\zeta) M_b(\zeta)^* x_b.$$

By (f1) we see that the right-hand side of this last equality is equal to

$$M(\zeta)^* B(\zeta) M_b(\zeta) x_b' + P_{H^2} \Omega_b(\zeta) M_b(\zeta)^* x_b.$$

We can write $\Omega_b(\zeta) = \Omega_{b0} + \Omega_{b1}\zeta + \cdots + \Omega_{bn}\zeta^n$. The fact that $x_b \in H(M_b)$ implies

$$M_b(\zeta)^* x_b = \zeta^{-1} u_{-1} + \zeta^{-2} u_{-2} + \cdots$$

for some $u_{-i} \in \mathbf{C}^m$, $i \geqq 1$. Therefore,

$$P_{H^2} \Omega_b(\zeta) M_b(\zeta)^* x_b = \sum_{i=1}^{n} \sum_{j=i}^{n} \Omega_{bj} \zeta^{j-i} u_{-i} =: x_{\omega b}.$$

Combining the above computations we get

$$P_{H(M)} B(S)x = B(S)x_b - M x_{\omega b}.$$

Similarly, for the computation of

$$P_{L(M)}C(U)y = C(U)y - MP_{H^2}M^*Cy,$$

we use the factorization

$$M(\zeta)^*C(\zeta) = \Omega_c(\zeta)M_c(\zeta)^*, \tag{f2}$$

where $\Omega_c(\zeta)$ is a polynomial matrix of size $p \times l$ and M_c is an inner matrix of size $l \times l$. As before we write $y = y_c + M_c y_c'$ where $y_c \in L(M)$ and $y_c' \in H^2(\mathbf{C}^l)$. Let $\Omega_c(\zeta) = \Omega_{c0} + \Omega_{c1}\zeta + \cdots + \Omega_{cn}\zeta^n$. Then

$$P_{H^2}\Omega_c M_c^* y_c = \sum_{i=1}^{n}\sum_{j=i}^{n} \Omega_{cj}\zeta^{j-i}v_{-i} =: y_{wc}$$

for some $v_{-i} \in \mathbf{C}^l$, $i = 1, \cdots, n$. This leads us to

$$P_{L(M)}C(U)y = C(U)y_c - My_{wc}.$$

Now we see that, with the above factorizations and decompositions, (1a), (1b) are equivalent to

$$\begin{aligned}&(\rho^2 k(S)^*k(S)I - D(S)^*D(S) - B(S)^*B(S))x_b\\ &-P_{H^2}((B(U)^*C(U) + D(U)^*E(U))y_c + D(U)^*E(U)M_c y_c')\\ &+(\rho^2 k(S)^*k(S)I - D(S)^*D(S))M_b x_b' = -B(S)^*Mx_{wb} - P_{H^2}B(U)^*My_{wc},\end{aligned} \tag{2a}$$

and

$$\begin{aligned}&(\rho^2 k(U)^*k(U)I - E(U)^*E(U) - C(U)^*C(U))y_c\\ &-(C(U)^*B(S) + E(U)^*D(S))x_b - E(U)^*D(S)M_b x_b'\\ &+(\rho^2 k(U)^*k(U)I - E(U)^*E(U))M_c y_c' = -C(U)^*My_{wc} - C(U)^*Mx_{wb}.\end{aligned} \tag{2b}$$

Now we will compute $P_{H^2}(B(U)^*C(U) + D(U)^*E(U))y_c$. First write

$$B(U)^*C(U) + D(U)^*E(U) = Q_{-n}^1 U^{*n} + \cdots + Q_0^1 + \cdots + Q_n^1 U^n.$$

Then,

$$P_{H^2}Q_{-i}^1 U^{*i}y_c = Q_{-i}^1 P_{H^2}U^{*i}y_c = Q_{-i}^1 S^{*i}(P_{H^2}y_c).$$

Let $y_c = \cdots + y_{c(-1)}\zeta^{-1} + y_{c(0)} + y_{c(1)}\zeta + \cdots$. Then

$$P_{H^2}Q_i^1 U^i y_c = Q_i^1 S^i(P_{H^2}y_c) + Q_i^1(\zeta^{i-1}y_{c(-1)} + \cdots + y_{c(-i)}).$$

Therefore,

$$\begin{aligned}&P_{H^2}(B(U)^*C(U) + D(U)^*E(U))y_c\\ &= (B(S)^*C(S) + D(S)^*E(S))(P_{H^2}y_c) + \sum_{i=1}^{n} Q_i^1(\zeta^{i-1}y_{c(-1)} + \cdots + y_{c(-i)}).\end{aligned}$$

Similarly, we have

$$P_{H^2}D(U)^*E(U)M_c y_c' = D(S)^*E(S)M_c y_c'.$$

Hence (2a) is equivalent to

$$\begin{aligned}&(\rho^2 k(S)^*k(S)I - D(S)^*D(S) - B(S)^*B(S))x_b + (\rho^2 k(S)^*k(S)I - D(S)^*D(S))M_b x_b'\\ &-((B(S)^*C(S) + D(S)^*E(S))y_c^+ + D(S)^*E(S)M_c y_c')\\ &= -B(S)^*Mx_{wb} - B(S)^*My_{wc} + \sum_{i=1}^{n} Q_i^1(\zeta^{i-1}y_{c(-1)} + \cdots + y_{c(-i)}),\end{aligned} \tag{3a}$$

where $y_c^+ := P_{H^2}y_c$. Note that we have $y = y_c^- + y_c^+ + M_c y_c'$, where $y_c^- \in L(M_c) \ominus H(M_c)$, $y_c^+ \in H(M_c)$, and $y_c' \in H^2(\mathbf{C}^l)$.

We will separate the equation (2b) into two parts by taking the orthogonal projections on $H^2(\mathbf{C}^l)$ and $L^2(\mathbf{C}^l)\ominus H^2(\mathbf{C}^l)$. As in the above discussion, if

$$\rho^2 k(U)^* k(U) I - E(U)^* E(U) - C(U)^* C(U) =: Q^2_{-n} U^{*n} + \cdots + Q^2_0 + \cdots + Q^2_n U^n,$$

then we have

$$P_{H^2}(\rho^2 k(U)^* k(U) I - E(U)^* E(U) - C(U)^* C(U)) y_c$$

$$= (\rho^2 k(S)^* k(S) I - E(S)^* E(S) - C(S)^* C(S)) y_c^+ + \sum_{i=1}^{n} Q_i^2(\zeta^{i-1} y_{c(-1)} + \cdots + y_{c(-i)}).$$

Hence the projection of (2b) on $H^2(\mathbf{C}^l)$ gives

$$(\rho^2 k(S)^* k(S) I - E(S)^* E(S) - C(S)^* C(S)) y_c^+ + (\rho^2 k(S)^* k(S) I - E(S)^* E(S)) M_c y_c'$$

$$-(C(S)^* B(S) + E(S)^* D(S)) x_b - E(S)^* D(S) M_b x_b' \tag{3b}$$

$$= -C(S)^* M y_{wc} - C(S)^* M x_{wb} - \sum_{i=1}^{n} Q_i^2(\zeta^{i-1} y_{c(-1)} + \cdots + y_{c(-i)}).$$

We now study the projection of (2b) on $L^2(\mathbf{C}^l)\ominus H^2(\mathbf{C}^l)$. First note that

$$P_{L^2\ominus H^2} Q^2_{-i} U^{*i} y_c = Q^2_{-i} U^{*i} y_c^- + Q^2_{-i}(\zeta^{-i} y_{c0} + \cdots + \zeta^{-1} y_{c(i-1)}),$$

and

$$P_{L^2\ominus H^2} Q_i^2 U^i y_c = Q_i^2 U^i y_c^- - Q_i^2(\zeta^{i-1} y_{c(-1)} + \cdots + y_{c(-i)}).$$

Hence

$$P_{L^2\ominus H^2}(\rho^2 k(U)^* k(U) I - E(U)^* E(U) - C(U)^* C(U)) y_c$$

$$= (\rho^2 k(U)^* k(U) I - E(U)^* E(U) - C(U)^* C(U)) y_c^-$$

$$+ \sum_{i=1}^{n} Q^2_{-i}(\zeta^{-i} y_{c0} + \cdots + \zeta^{-1} y_{c(i-1)})$$

$$- \sum_{i=1}^{n} Q_i^2(\zeta^{i-1} y_{c(-1)} + \cdots + y_{c(-i)}).$$

This takes care of the first term in (2b). For the projections of the other terms we use the following notation:

$$\rho^2 k(U)^* k(U) I - E(U)^* E(U) =: Q^3_{-n} U^{*n} + \cdots + Q^3_0 + \cdots + Q^3_n U^n,$$

$$C(U)^* B(U) + E(U)^* D(U) =: Q^4_{-n} U^{*n} + \cdots + Q^4_0 + \cdots + Q^4_n U^n,$$

$$E(U)^* D(U) =: Q^5_{-n} U^{*n} + \cdots + Q^5_0 + \cdots + Q^5_n U^n,$$

$$M_b(\zeta) =: M_{b0} + M_{b1}\zeta^1 + M_{b2}\zeta^2 + \cdots,$$

$$M_c(\zeta) =: M_{c0} + M_{c1}\zeta^1 + M_{c2}\zeta^2 + \cdots,$$

$$M(\zeta) =: M_0 + M_1\zeta^1 + M_2\zeta^2 + \cdots,$$

$$C(U)^* =: C_0^* + C_1^* U^* + \cdots + C_n^* U^{*n},$$

$$x_b'(\zeta) =: x_{b0}' + x_{b1}'\zeta^1 + \cdots,$$

$$y_c'(\zeta) =: y_{c0}' + y_{c1}'\zeta^1 + \cdots,$$

$$x_b(\zeta) =: x_{b0} + x_{b1}\zeta^1 + \cdots.$$

With this notation, taking the projection of (2b) on $L^2(\mathbf{C}^l)\ominus H^2(\mathbf{C}^l)$, and then multiplying both sides of the resulting equation by ζ^{-n} (this is equivalent to the operation U^{*n}, which is left invertible on $L^2(\mathbf{C}^l)\ominus H^2(\mathbf{C}^l)$) will give us

$$X_3(\zeta^{-1})y_c^- := F_3(\zeta^{-1}), \tag{4a}$$

where $X_3(\zeta^{-1}) = Q^2_{-n}\zeta^{-2n} + \cdots + Q^2_0\zeta^{-n} + \cdots + Q^2_n$, and

$$\begin{aligned} F_3(\zeta^{-1}) := {} & \sum_{i=1}^{n} Q_i^2 \sum_{j=1}^{i} \zeta^{-n+i-j} y_{c(-j)} - \sum_{i=1}^{n} Q^2_{-i} \sum_{j=0}^{i-1} \zeta^{-n+i-j} y_{c(j)} \\ & - \sum_{i=1}^{n} C_i^* \sum_{j=0}^{i-1} \zeta^{-n+j-i} \sum_{k=0}^{j} M_{j-k} \sum_{s=1}^{n-k} \Omega_{c(s+k)} v_{-s} \\ & + \sum_{i=1}^{n} Q^4_{-i} \sum_{j=0}^{i-1} \zeta^{-n-i+j} x_{bj} + \sum_{i=1}^{n} Q^5_{-i} \sum_{j=1}^{i} \zeta^{-n-j} \sum_{k=0}^{i-j} M_{b(i-j-k)} x'_{bk} \\ & - \sum_{i=0}^{n} Q^3_{-i} \sum_{j=1}^{i} \zeta^{-n-j} \sum_{k=0}^{i-j} M_{c(i-j-k)} y'_{ck} \\ & - \sum_{i=1}^{n} C_i^* \sum_{j=0}^{i-1} \zeta^{-n+j-i} \sum_{k=0}^{j} M_{j-k} \sum_{s=1}^{n-k} \Omega_{b(s+k)} u_{-s}. \end{aligned}$$

We now play the same game with (3a) and (3b). Indeed, we multiply both sides of these equations by ζ^n. (This is equivalent to the application of the operator S^n, which is left invertible on $H^2(\mathbf{C}^l)$ and $H^2(\mathbf{C}^m)$.) Set

$$\begin{aligned} \rho^2 k(S)^* k(S) I - D(S)^* D(S) - B(S)^* B(S) &=: Q^6_{-n} S^{*n} + \cdots + Q^6_0 + \cdots + Q^6_n S^n, \\ \rho^2 k(S)^* k(S) I - D(S)^* D(S) &=: Q^7_{-n} S^{*n} + \cdots + Q^7_0 + \cdots + Q^7_n S^n, \\ D(S)^* E(S) &=: Q^8_{-n} S^{*n} + \cdots + Q^8_0 + \cdots + Q^8_n S^n, \\ B(S)^* &=: B_0^* + \cdots + B_n^* S^{*n}. \end{aligned}$$

For any polynomial of degree $\leqq n$, $P(\zeta)$, we define $\hat{P}(\zeta) := \zeta^n P^*(\zeta^{-1})$. Then it is easy to see that (3a) combined with (3b) is equivalent to

$$X_1(\zeta)\begin{bmatrix} x_b \\ y_c^+ \end{bmatrix} + X_2(\zeta)\begin{bmatrix} M_b & 0 \\ 0 & M_c \end{bmatrix}\begin{bmatrix} x'_b \\ y'_c \end{bmatrix} = \begin{bmatrix} F_1(\zeta) \\ F_2(\zeta) \end{bmatrix}, \tag{4b}$$

where

$$X_1(\zeta) := \begin{bmatrix} (\rho^2 \hat{k}kI - \hat{D}D - \hat{B}B) & -(\hat{B}C + \hat{D}E) \\ -(\hat{C}B + \hat{E}D) & (\rho^2 \hat{k}kI - \hat{E}E - \hat{C}C) \end{bmatrix},$$

$$X_2(\zeta) := \begin{bmatrix} (\rho^2 \hat{k}kI - \hat{D}D) & -\hat{D}E \\ -\hat{E}D & (\rho^2 \hat{k}kI - \hat{E}E) \end{bmatrix},$$

$$\begin{aligned} F_1(\zeta) := {} & \sum_{i=1}^{n} Q^6_{-i} \sum_{j=0}^{i-1} \zeta^{n-i+j} x_{bj} + \sum_{i=1}^{n} Q^7_{-i} \sum_{j=0}^{i-1} \zeta^{n-i+j} \sum_{k=0}^{j} M_{b(j-k)} x'_{bk} \\ & + \sum_{i=1}^{n} Q_i^1 \sum_{j=1}^{i} \zeta^{n+i-j} y_{c(-j)} - \sum_{i=1}^{n} Q^1_{-i} \sum_{j=0}^{i-1} \zeta^{n-i+j} y_{cj} \\ & - \sum_{i=1}^{n} Q^8_{-i} \sum_{j=0}^{i-1} \zeta^{n-i+j} \sum_{k=0}^{j} M_{c(j-k)} y'_{ck} - \hat{B}(\zeta) M(\zeta) \sum_{i=1}^{n} \sum_{j=1}^{i} \Omega_{bi} \zeta^{i-j} u_{-j} \\ & + \sum_{i=1}^{n} B_i^* \sum_{j=0}^{i-1} \zeta^{n-i+j} \sum_{k=0}^{j} M_{j-k} \sum_{s=1}^{n-k} \Omega_{b(s+k)} u_{-s} - \hat{B}(\zeta) M(\zeta) \sum_{i=1}^{n} \sum_{j=1}^{i} \Omega_{ci} \zeta^{i-j} v_{-j} \\ & + \sum_{i=1}^{n} B_i^* \sum_{j=0}^{i-1} \zeta^{n-i+j} \sum_{k=0}^{j} M_{j-k} \sum_{s=1}^{n-k} \Omega_{c(s+k)} v_{-s}, \end{aligned}$$

and

$$F_2(\zeta) = \sum_{i=1}^{n} Q_{-i}^2 \sum_{j=0}^{i-1} \zeta^{n-i+j} y_{cj} + \sum_{i=1}^{n} Q_{-i}^3 \sum_{j=0}^{i-1} \zeta^{n-i+j} \sum_{k=0}^{j} M_{c(j-k)} y'_{ck}$$

$$- \sum_{i=1}^{n} Q_i^2 \sum_{j=1}^{i} \zeta^{n+i-j} y_{c(-j)} - \sum_{i=1}^{n} Q_{-i}^4 \sum_{j=0}^{i-1} \zeta^{n-i+j} x_{bj}$$

$$- \sum_{i=1}^{n} Q_{-i}^5 \sum_{j=0}^{i-1} \zeta^{n-i+j} \sum_{k=0}^{j} M_{b(j-k)} x'_{bk} - \hat{C}(\zeta) M(\zeta) \sum_{i=1}^{n} \sum_{j=1}^{i} \Omega_{bi} \zeta^{i-j} u_{-j}$$

$$+ \sum_{i=1}^{n} C_i^* \sum_{j=0}^{i-1} \zeta^{n-i+j} \sum_{k=0}^{j} M_{j-k} \sum_{s=1}^{n-k} \Omega_{b(s+k)} u_{-s} - \hat{C}(\zeta) M(\zeta) \sum_{i=1}^{n} \sum_{j=1}^{i} \Omega_{ci} \zeta^{i-j} v_{-j}$$

$$+ \sum_{i=1}^{n} C_i^* \sum_{j=0}^{i-1} \zeta^{n-i+j} \sum_{k=0}^{j} M_{j-k} \sum_{s=1}^{n-k} \Omega_{c(s+k)} v_{-s}.$$

Let us summarize the above results in the following:

PROPOSITION 1. *ρ^2 is an eigenvalue of A^*A if and only if there exists $x_b \in H(M_b)$, $x'_b \in H^2(\mathbf{C}^m)$, $y_c^+ \in H(M_c)$, $y_c^- \in L(M_c) \ominus H(M_c)$, $y'_2 \in H^2(\mathbf{C}^l)$, not all zero, such that* (4a) *and* (4b) *hold.*

Defining

$$M_0 := \begin{bmatrix} M_b & 0 \\ 0 & M_c \end{bmatrix},$$

we see that

$$\begin{bmatrix} x_b \\ y_c^+ \end{bmatrix} \in H(M_0) = H^2(\mathbf{C}^N) \ominus M_0 H^2(\mathbf{C}^N), \qquad N = m + l.$$

Now set

$$x_0 := \begin{bmatrix} x_b \\ y_c^+ \end{bmatrix}, \quad x'_0 := \begin{bmatrix} x'_b \\ y'_c \end{bmatrix}, \quad F'_0 := \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \cdot p.$$

Then (4b) can be rewritten as

$$X_1(\zeta) x_0 + X_2(\zeta) M_0 x'_0 = F'_0(\zeta). \tag{5}$$

Remark. Equation (5) is exactly the same type of equation that we obtained in [12] for the 2-block problem. In the 1-block case, we get a similar equation with $X_2(\zeta)$, a scalar. In fact, if we assume that $d_b(\zeta) := \det X_2(\zeta)$ is not identically equal to zero, then (5) can be put in the form

$$X_0(\zeta) x_0 + d_b(\zeta) M_0 x'_0 = F_0(\zeta), \tag{6}$$

where $X_0 = X_2^a X_1$, $F_0 = X_2^a F'_0$, and $X_2^a(\zeta)$ is the algebraic adjoint of $X_2(\zeta)$, i.e.,

$$X_2^a(\zeta) X_2(\zeta) = X_2(\zeta) X_2^a(\zeta) = d_b(\zeta) I.$$

For (6), we make the factorization

$$X_0(\zeta) M_1(\zeta) = M_0(\zeta) \Omega_0(\zeta), \tag{f3}$$

where $M_1(\zeta)$ is $N \times N$ inner and $\Omega_0(\zeta)$ is $N \times N$ polynomial. Then, as shown using skew Toeplitz theory in [3], there exists $X_0^{(-1)}$, an $N \times N$ H^∞-matrix, such that

$$X_0^{(-1)} X_0 = I + M_1 E_0 \quad \text{and} \quad X_0^{(-1)} M_0 = M_1 E_1$$

for some E_0 and E_1, $N \times N$ H^∞-matrices. Multiplying both sides of (6) by $X_0^{(-1)}$ and taking the orthogonal projection, of the resulting equation, on $H(M_1)$ we obtain

$$P_{H(M_1)} x_0 = P_{H(M_1)} X_0^{(-1)} F_0(\zeta).$$

Now we make our first assumption of genericity.

Assumption (a1). The operator $\tau := P_{H(M_1)}|_{H(M_0)}$ is invertible.

With this assumption we obtain

$$x_0 = \tau^{-1} P_{H(M_1)} X_0^{(-1)} F_0, \tag{7a}$$

and

$$d_b(\zeta) x_0' = P_{H^2}(M_0(\zeta)^*(I - X_0 \tau^{-1} P_{H(M_1)} X_0^{(-1)}) F_0). \tag{7b}$$

Next, applying the algebraic adjoint of $X_3(\zeta^{-1})$, $X_3^a(\zeta^{-1})$ to both sides of (4a) we get

$$d_d(\zeta^{-1}) y_c^- = X_3^a(\zeta^{-1}) F_3(\zeta^{-1}), \tag{7c}$$

where $d_d(\zeta^{-1}) = \det X_3(\zeta^{-1})$. Equation (7a) gives the conditions for invertibility of a certain skew Toeplitz operator. See also [3]. We see that it is coupled, via F_0, to (7b) and (7c), which give the invertibility conditions of two Toeplitz operators.

We will now show that (7a)-(7c) give finitely many interpolation conditions for ρ^2 to be an eigenvalue of A^*A. From this we will derive the finite matricial rank condition for the determination of the singular values of A. First note that there exists $y_c^- \in L(M_c) \ominus H(M_c)$ satisfying (7c) if and only if there exists $\hat{y}_c^- \in H^2(\mathbf{C}^l)$ satisfying

$$d_d(\zeta) \hat{y}_c^- = P_{H^2}(\zeta^{-1} X_3^a(\zeta) F_3(\zeta)). \tag{7d}$$

Indeed, this follows since $L^2(\mathbf{C}^l) \ominus H^2(\mathbf{C}^l)$ is isomorphic to $SH^2(\mathbf{C}^l) = \zeta H^2(\mathbf{C}^l)$ and the natural isomorphism is given by the reflection operator: $\zeta^{-1} \to \zeta$.

Next it is easy to see that the right-hand sides of (7a), (7b), and (7d) can be put into the form

$$\tau^{-1} P_{H(M_1)} X_0^{(-1)}(\zeta) F_0(\zeta) = K_a(\zeta)\Phi,$$

$$P_{H^2}(M_0(\zeta)^*(I - X_0(\zeta)\tau^{-1} P_{H(M_1)} X_0^{(-1)}(\zeta)) F_0(\zeta)) = K_b(\zeta)\Phi,$$

$$P_{H^2}(\zeta^{-1} X_3^a(\zeta) F_3(\zeta)) = K_d(\zeta)\Phi,$$

where $K_a(\zeta)$, $K_b(\zeta)$ are H^∞ matrices of sizes $N \times r$ and $K_d(\zeta)$ is an $l \times r$ polynomial in ζ (these all can be explicitly computed from M_0, M_1, X_0, $X_0^{(-1)}$, X_3, F_0, and F_3),

$$\Phi^T = [x_{b0}^T \cdots x_{b(n-1)}^T x_{b0}'^T \cdots x_{b(n-1)}'^T y_{c0}^T \cdots y_{c(n-1)}^T y_{c0}'^T$$
$$\cdots y_{c(n-1)}'^T u_{-1}^T \cdots u_{-n}^T v_{-1}^T \cdots v_{-n}^T y_{c(-1)}^T \cdots y_{c(-n)}^T],$$

and $r = 2n(m+l) + n(m+2l)$. With this notation we immediately get the following identities:

$$K_{ai}\Phi = x_{0i}, \tag{8a}$$

$$K_{bi}\Phi = \sum_{j=0}^{i} d_{bj} x_{0(i-j)}', \tag{8b}$$

$$K_{di}\Phi = \sum_{j=0}^{i} d_{dj} \hat{y}_{c(i-j)}^-, \tag{8d}$$

for all $i=0,\cdots,n$, where

$$K_a(\zeta)=:K_{a0}+K_{a1}\zeta+K_{a2}\zeta^2+\cdots,$$
$$K_b(\zeta)=:K_{b0}+K_{b1}\zeta+K_{b2}\zeta^2+\cdots,$$
$$K_d(\zeta)=:K_{d0}+K_{d1}\zeta+K_{d2}\zeta^2+\cdots,$$
$$x_0(\zeta)=:x_{00}+x_{01}\zeta+x_{02}\zeta^2+\cdots,$$
$$x_0'(\zeta)=:x_{00}'+x_{01}'\zeta+x_{02}'+\cdots,$$
$$\hat{y}_c^-(\zeta)=:\hat{y}_{c0}^-+\hat{y}_{c1}^-\zeta+\hat{y}_{c2}^-\zeta^2+\cdots,$$
$$d_b(\zeta)=:d_{b0}+\cdots+d_{b2nN}\zeta^{2nN},$$
$$d_d(\zeta)=:d_{d0}+\cdots+d_{d2nl}\zeta^{2nl}.$$

Rearranging terms in (8a), (8b), (8d) and combining them into one equation we obtain,

$$K\Phi=0, \tag{9}$$

where K is a constant matrix that can be computed from the K_{ai}, K_{bi}, K_{di}, d_{bi}, and d_{di}, $i=0,\cdots,n$.

We now make our second assumption of genericity.

Assumption (a2). $d_b(\zeta)$ and $d_d(\zeta)$ have distinct roots, all of which are nonzero.

Then, as in [6], [7], and [10], we see that d_b has roots $\alpha_1,\cdots,\alpha_{r_b}$ inside D, $\alpha_{r_b+1},\cdots,\alpha_{(2nN-r_b)}$ on ∂D and $1/\bar{\alpha}_1,\cdots,1/\bar{\alpha}_{r_b}$ outside $\bar{D}$. Similarly, d_d has roots $\beta_1,\cdots,\beta_{r_d}$ inside D, $\beta_{r_d+1},\cdots,\beta_{(2nl-r_d)}$ on ∂D and $1/\bar{\beta}_1,\cdots,1/\bar{\beta}_{r_d}$ outside $\bar{D}$.

We are ready to state our main result.

THEOREM 1. *Assume* (a1) *and* (a2). *Then,* $\rho^2>\|A\|_e^2$ *is an eigenvalue of* A^*A *if and only if*

$$\operatorname{rank} R<r,$$

where

$$R:=\begin{bmatrix} K \\ K_b(\alpha_1) \\ \vdots \\ K_b(\alpha_{(2nN-r_b)}) \\ K_d(\beta_1) \\ \vdots \\ K_d(\beta_{(2nl-r_d)}) \end{bmatrix}. \tag{9a}$$

Proof. By Proposition 1, ρ^2 is an eigenvalue of A^*A if and only if there exists $x_0\in H(M_0)$, $x_0'\in H^2(\mathbf{C}^N)$ and $\hat{y}_c^-\in H^2(\mathbf{C}^l)$, not zero, such that (7a), (7b), (7d) are satisfied. By an argument similar to the one used in [3], [6], [7], and [11], we see that the existence of such x_0, x_0', $\hat{y}_c^-$ is equivalent to finding a nonzero Φ such that

$$K_b(\alpha_i)\Phi=0, \qquad i=1,\cdots,2nN-r_b,$$
$$K_d(\beta_i)\Phi=0, \qquad i=1,\cdots,2nl-r_d,$$

and (9) holds. This completes the proof. □

Remark. In the absence of the genericity assumptions, the matrix (9a) takes on a certain degenerate form exactly as in [11]. We see from Theorem 1 that the largest value of ρ that gives a solution for the equation

$$\det R^*R=0$$

is the norm of the 4-block operator A. From R, we can determine the singular values and singular vectors of the 4-block operator A.

3.2. Essential spectrum. We now give a sufficient condition for ρ to be strictly greater than the essential norm of A; in order to do this we study the essential spectrum of A^*A.

PROPOSITION 2. *Suppose that*

(a3) *the Toeplitz operator* $\tau_2 := P_{H^2} M_1^* M_0|_{H^2}$ *is invertible,*

(a4) $\{z: \det X_1(z) = 0\} \cap \sigma_e(T_0) = \varnothing$,

where $T_0 := P_{H(M_0)} S|_{H(M_0)}$. *Then,* $\rho > \max\{\beta, \gamma\}$ *implies* $\rho^2 \notin \sigma_e(A^*A)$, *where* β *and* γ *are defined as in* § 2.

Proof. Let $\rho > \max\{\beta, \gamma\}$. If ρ^2 were in $\sigma_e(A^*A)$, then there would exist

$$\begin{bmatrix} x^{(n)} \\ y^{(n)} \end{bmatrix} \in H^2(\mathbf{C}^m) \oplus L^2(\mathbf{C}^l) \quad \text{with} \quad \left\| \begin{bmatrix} x^{(n)} \\ y^{(n)} \end{bmatrix} \right\|_2 = 1 \quad \forall n \geq 1,$$

and $\left[\begin{smallmatrix} x^{(n)} \\ y^{(n)} \end{smallmatrix}\right] \to 0$ weakly as $n \to \infty$, satisfying

$$(4b)_e \qquad X_1(\zeta) \begin{bmatrix} x_b^{(n)} \\ y_c^{+(n)} \end{bmatrix} + X_2(\zeta) \begin{bmatrix} M_b & 0 \\ 0 & M_c \end{bmatrix} \begin{bmatrix} x_b'^{(n)} \\ y_c'^{(n)} \end{bmatrix} \to 0 \quad \text{strongly,}$$

and

$$(7d)_e \qquad X_3(\zeta) \hat{y}_c^{-(n)} \to 0 \quad \text{strongly.}$$

(These conditions for $\rho^2 \in \sigma_e(A^*A)$ are sufficient as well.) This follows from Proposition 1 and equations (4b) and (7d). Note that $F_1(\zeta)$, $F_2(\zeta)$, and $F_3(\zeta)$ converge to zero strongly as $x^{(n)}$ and $y^{(n)}$ converge to zero weakly. In the above we have, as before,

$$x^{(n)} = x_b^{(n)} + M_b x_b'^{(n)},$$

$$y^{(n)} = y_c^{-(n)} + y_c^{+(n)} + M_c y_c'^{(n)},$$

with

$$\hat{y}_c^{-(n)}(\zeta) := y_c^{-(n)}(\zeta^{-1}),$$

$$\begin{bmatrix} x_b^{(n)} \\ y_c^{+(n)} \end{bmatrix} =: x_0^{(n)} \in H(M_0),$$

$$\begin{bmatrix} x_b'^{(n)} \\ y_c'^{(n)} \end{bmatrix} =: x_0'^{(n)} \in H^2(\mathbf{C}^m) \oplus H^2(\mathbf{C}^l),$$

and $y_c^{-(n)} \in L^2(\mathbf{C}^l) \ominus H^2(\mathbf{C}^l)$. They all converge to zero weakly as $n \to \infty$.

Note that $(7d)_e$ means that

$$\left(\rho^2 I - [F(S)^* \quad J(S)^*] \begin{bmatrix} F(S) \\ J(S) \end{bmatrix} \right) \hat{y}_c^{-(n)} \to 0 \quad \text{strongly.}$$

Since $\rho > \gamma$, we see that

$$\left(\rho^2 I - [F(S)^* \quad J(S)^*] \begin{bmatrix} F(S) \\ J(S) \end{bmatrix} \right)$$

is invertible, and so $\hat{y}_c^{-(n)}$ converges to zero strongly.

Next from $(4b)_e$ we get that

$$(6)_e \qquad X_0(\zeta) x_0^{(n)} + d_b(\zeta) M_0 x_0'^{(n)} \to 0 \quad \text{strongly.}$$

Taking orthogonal projections on $M_1 H^2(\mathbf{C}^N)$ we see that

$$P_{H^2} M_1^* X_0 x_0^{(n)} + P_{H^2} M_1^* M_0 d_b(\zeta) x_0'^{(n)} \to 0 \quad \text{strongly.}$$

Recall that $P_{H^2} M_1^* X_0 x_0^{(n)} = P_{H^2} \Omega_0 M_0^* x_0^{(n)}$, so it converges to zero strongly as $x_0^{(n)} \in H(M_0)$ converges to zero weakly. Hence using Assumption (a3) we have that

$$(7)_e \qquad d_b(\zeta) x_0'^{(n)} \to 0 \quad \text{strongly.}$$

This implies, by $(6)_e$, that

$$(8)_e \qquad d_b(\zeta) \det X_1(\zeta) x_0^{(n)} \to 0 \quad \text{strongly.}$$

It is easy to see, by definition of β, that for $\rho > \beta$, $d_b(\zeta)$ has no roots on ∂D. Then we can write

$$d_b(\zeta) = (\zeta - \alpha_1)\left(\zeta - \frac{1}{\bar{\alpha}_1}\right) \cdots (\zeta - \alpha_{nN})\left(\zeta - \frac{1}{\bar{\alpha}_{nN}}\right),$$

for some $\alpha_1, \cdots, \alpha_{nN} \in D$. Multiplying $(7)_e$ by

$$\prod_{i=1}^{nN} \frac{\bar{\alpha}_i}{(1 - \zeta \bar{\alpha}_i)^2},$$

which is in H^∞ (because all α_i's are in D), we obtain

$$m_1(\zeta) x_0'^{(n)} \to 0 \quad \text{strongly,}$$

where

$$m_1(\zeta) = \prod_{i=1}^{nN} \frac{\zeta - \alpha_i}{(1 - \zeta \bar{\alpha}_i)}.$$

This implies that $x_0'^{(n)} \to 0$ strongly, because $m_1(S)^* m_1(S)$ is equal to the identity.

From $(8)_e$, a similar argument gives that

$$(9)_e \qquad \det X_1(\zeta) x_0^{(n)} \to 0 \quad \text{strongly.}$$

Let us assume now that $d_1(\zeta) := \det X_1(\zeta)$ has nonzero distinct roots. So $d_1(\zeta) = 0$ at points $z_1, \cdots, z_{n_1}$ inside D, $1/\bar{z}_1, \cdots, 1/\bar{z}_{n_1}$ outside $\bar{D}$, and $z_{n_1+1}, \bar{z}_{n_1+1}, \cdots, z_{nN}, \bar{z}_{nN}$ on ∂D. Using a similar trick as before, we obtain

$$m_2(\zeta) \prod_{i=n_1+1}^{nN} (\zeta - z_i)(\zeta - \bar{z}_i) x_0^{(n)} \to 0 \quad \text{strongly,}$$

where

$$m_2(\zeta) = \prod_{i=1}^{n_1} \frac{\zeta - z_i}{1 - \bar{z}_i \zeta}.$$

Hence we see that

$$\prod_{i=n_1+1}^{nN} (\zeta - z_i)(\zeta - \bar{z}_i) x_0^{(n)} \to 0 \quad \text{strongly.}$$

Taking the orthogonal projection of this last expression on $H(M_0)$, we get

$$\prod_{i=n_1+1}^{nN} (T_0 - z_i)(T_0 - \bar{z}_i) x_0^{(n)} \to 0 \quad \text{strongly,}$$

for $x_0^{(n)} \to 0$ weakly, and $x_0^{(n)} \in H(M_0)$. By Assumption (a4) none of z_i, $\bar{z}_i$ are in the essential spectrum of T_0, therefore $x_0^{(n)} \to 0$ strongly.

In summary, we have established that $\hat{y}_c^{-(n)} \to 0$ strongly, $x_0'^{(n)} \to 0$ strongly and $x_0^{(n)} \to 0$ strongly. So, $\begin{bmatrix} x_0^{(n)} \\ y^{(n)} \end{bmatrix} \to 0$ strongly, which contradicts that $\left\| \begin{bmatrix} x_0^{(n)} \\ y^{(n)} \end{bmatrix} \right\|_2 = 1$ for all $n \geqq 1$. Thus ρ^2 cannot be in $\sigma_e(A^*A)$. □

Remark. Note that a sufficient condition for (a4) to hold is $\rho > \alpha$. So if $\|A\|_e > \alpha$, then the above Proposition 2 gives an upper bound for the essential norm: $\|A\|_e \leqq \max\{\beta, \gamma\}$. Actually, we must prove, if possible, the equality as in the case of SISO plants and MIMO finite-dimensional systems. However this is not easy in our case: All we can show is that if $\rho = \gamma$ then $(7d)_e$ holds for some $\hat{y}_c^{-(n)}$ such that $\|\hat{y}_c^{-(n)}\|_2 = 1$ for all $n \geqq 1$, and $\hat{y}_c^{-(n)} \to 0$ weakly. This implies that $\rho^2 \in \sigma_e(A^*A)$, and $\|A\|_e \geqq \gamma$. But the difficulty is with β: if $\rho = \beta$ then there exists

$$\begin{bmatrix} x^{(n)} \\ y^{(n)} \end{bmatrix} \in H^2(\mathbf{C}^m) \oplus H^2(\mathbf{C}^l) \quad \text{with} \quad \left\| \begin{bmatrix} x^{(n)} \\ y^{(n)} \end{bmatrix} \right\|_2 = 1 \quad \forall n \geqq 1,$$

and $\begin{bmatrix} x^{(n)} \\ y^{(n)} \end{bmatrix} \to 0$ weakly as $n \to \infty$, such that

$$(10)_e \qquad X_2(\zeta) \begin{bmatrix} x^{(n)} \\ y^{(n)} \end{bmatrix} \to 0 \quad \text{strongly.}$$

In the SISO case, by multiplying $(10)_e$ by $M_0(\zeta)$ (which then commutes with $X_2(\zeta)$), we get the result that $\|A\|_e \geqq \beta$. Moreover, in the MIMO finite-dimensional case we decompose $\begin{bmatrix} x_0^{(n)} \\ y^{(n)} \end{bmatrix}$ as $x_0^{(n)} + M_0 x_0'^{(n)}$, and as before $x_0^{(n)} \in H(M_0)$. Since in the finite-dimensional case $H(M_0)$ is finite-dimensional, $x_0^{(n)} \to 0$ weakly implies $x_0^{(n)} \to 0$ strongly. Hence we obtain that $(4b)_e$ holds; then $\|A\|_e \geqq \beta$. The infinite-dimensional MIMO case is much more subtle.

We now summarize the above discussion with two corollaries to Proposition 2.

COROLLARY 1. *Assume* (a3) *and* $\alpha \leqq \max\{\beta, \gamma\}$. *Then,*

(i) *If* $\gamma \geqq \beta$ *then* $\|A\|_e = \gamma$.

(ii) *If* $\gamma < \beta$ *then* $\gamma \leqq \|A\|_e \leqq \beta$.

COROLLARY 2. *Consider finite-dimensional* MIMO *case, i.e.,* $M(\zeta)$ *is rational. Then,*

$$\|A\|_e = \max\{\beta, \gamma\}.$$

Proof. Let M^{ad} denote the algebraic adjoint of M. Then

$$\begin{bmatrix} M^{\mathrm{ad}} & 0 \\ 0 & M^{\mathrm{ad}} \end{bmatrix} \begin{bmatrix} W - MQ & F \\ G & J \end{bmatrix} = \begin{bmatrix} M^{\mathrm{ad}}W - mQ & M^{\mathrm{ad}}F \\ M^{\mathrm{ad}}G & M^{\mathrm{ad}}J \end{bmatrix} =: L,$$

where $m := \det M$. Clearly, L has all rational entries. Now let A_L be the 4-block operator associated to L. Then it is easy to see that $\|A_L\|_e = \|A\|_e$. In other words, without loss of generality, we may assume that M is of the form mI where $m \in H^\infty$ is an inner scalar-valued function. But in this case, we have that (a3) is satisfied since we can choose $M_1 = m$ (see also the discussion below in § 4). Hence by Proposition 2, and by the finite dimensionality of $H(m)$, we have the required conclusion. □

Remark. In practice we do not need to compute the essential norm. All we need to know is an upper bound μ_0 for $\|A\|$ with which to start. Then the first zero of $\det R^*R$ (considered as a function of ρ) less than μ_0, will be $\|A\|$. Of course, if $\|A\| = \|A\|_e$, then there is no first eigenvalue. Hence on the computer, if we plot $\det R^*R$ as a function of ρ, the graph of $\det R^*R$ does not cross the ρ axis above $\|A\|_e$, but oscillates near this value, since the eigenvalues accumulate at $\|A\|_e$. In this way we can estimate the essential norm.

4. SISO case. In this section we apply the above theory to SISO plants. The first thing to note in this case is that the factorizations (f1), (f2), (f3) are trivial, because $M(\zeta)$ is scalar, so it commutes with everything:

$$M(\zeta)^*B(\zeta) = B(\zeta)M(\zeta)^*,$$

$$M(\zeta)^*C(\zeta) = C(\zeta)M(\zeta)^*,$$

$$X_0(\zeta)M_0(\zeta) = M_0(\zeta)X_0(\zeta).$$

Here we have $M_0(\zeta) = M(\zeta)\left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]$. Since $M_0(\zeta) = M_1(\zeta)$ Assumption (a1) holds (in fact, τ is the identity). Moreover, we do not really have to compute $X_0^{(-1)}$. Indeed, recall the equation

$$X_1(\zeta)x_0 + X_2(\zeta)M_0(\zeta)x_0' = F_0'(\zeta). \tag{5}$$

Taking the projections of (5) on $H(M_0)$ and $M_0H^2(\mathbf{C}^N)$, we obtain

$$X_1(\zeta)x_0 = F_0'(\zeta) - M_0(\zeta)P_{H^2}M_0(\zeta)^*F_0'(\zeta) + M_0(\zeta)P_{H^2}M_0(\zeta)^*X_1(\zeta)x_0, \tag{5a}$$

and

$$X_2(\zeta)x_0' = P_{H^2}M_0(\zeta)^*F_0'(\zeta) - P_{H^2}M_0(\zeta)^*X_1(\zeta)x_0. \tag{5b}$$

These equations (5a), (5b), are in the form of the equations (24c), (24d) of [6]. Now we can use similar computations to the ones used in [6] to obtain the final result, namely, a rank type formula as in our main theorem.

In the next section we give an example illustrating the computations for the SISO case.

5. A SISO 2-block example. For simplicity of notation and exposition, the following example is chosen in the 2-block setup and a SISO plant is considered. The 2-block problem for stable SISO distributed plants was first solved in [22]. Motivations for studying the 2-block problem comes from the mixed sensitivity minimization (see, e.g., [14], [19]), which can be stated as follows. Consider the feedback configuration shown in Fig. 2. The mixed sensitivity minimization problem is to find

$$\mu = \inf_{C\text{stabilizing}} \sup\left\{\left\|\begin{bmatrix}\tilde{e}\\ \tilde{u}\end{bmatrix}\right\|_2 : \|v\|_2 \leqq 1\right\}$$

$$= \inf_{C\text{stabilizing}} \left\|\begin{bmatrix}W_1(I+PC)^{-1}W_3\\ W_2C(I+PC)^{-1}W_3\end{bmatrix}\right\|_\infty.$$

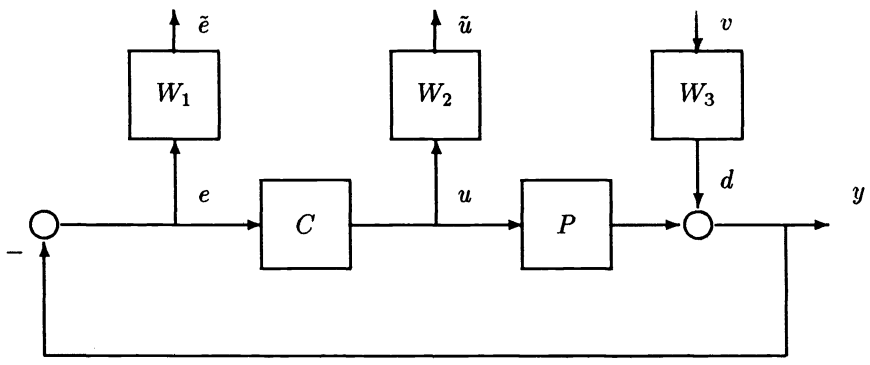


FIG. 2

Invoking the standard Youla parametrization of all stabilizing controllers, we obtain the following expression for μ for P stable:

$$\mu = \inf_{Z \in H^\infty} \left\| \begin{bmatrix} W_1(1-P)W_3 \\ W_2 W_3 \end{bmatrix} - \begin{bmatrix} W_1 P W_3 \\ -W_2 W_3 \end{bmatrix} Z \right\|_\infty .$$

Let us now choose some specific values for the weights and the plant: $W_1 = 1$, $W_2 = b$, $W_3 = 1/(s+1)$, and $P = e^{-hs}$. Here $0 \leqq b < \infty$ and $0 \leqq h < \infty$ are free parameters. We will find the dependence of μ on b and h. Note that if $b = 0$ then the problem reduces to the 1-block case.

Following the factorization techniques used in [15] and [19] we can show that

$$\mu = \inf_{Q \in H^\infty} \left\| \begin{bmatrix} \dfrac{1}{\sqrt{1+b^2}} & \dfrac{1}{s+1} \\ \dfrac{b}{\sqrt{1+b^2}} & \dfrac{1}{s+1} \end{bmatrix} - \begin{bmatrix} e^{-hs} \\ 0 \end{bmatrix} Q \right\|_\infty .$$

In terms of our notation

$$W(\zeta) = \frac{1}{\sqrt{1+b^2}} \frac{(1-\zeta)}{2}, \qquad G(\zeta) = \frac{b}{\sqrt{1+b^2}} \frac{(1-\zeta)}{2},$$

and $M(\zeta) = e^{h(\zeta+1)/(\zeta-1)}$. We can compute the lower bound for μ as $\|A\|_e = b/\sqrt{1+b^2}$. Also note that if we set $Q = 0$ then we find an upper bound for μ as one. Therefore we seek solutions ρ^2, to the eigenvalue equations (1a), (1b) in the region:

$$\frac{b^2}{1+b^2} \leqq \rho^2 \leqq 1.$$

In this specific example, equation (5) turns out to be

$$X_1(\zeta) x_b + X_2(\zeta) M(\zeta) x_b' = F_1(\zeta), \tag{11}$$

where $x_b \in H(M)$, $x_b' \in H^2$, and where X_1, X_2, and F_1 can be computed to be

$$X_1(\zeta) = \frac{1}{4}(\zeta^2 + (4\rho^2 - 2)\zeta + 1), \tag{11a}$$

$$X_2(\zeta) = \frac{b^2}{4(1+b^2)} \left(\zeta^2 + \left(\frac{4(1+b^2)}{b^2} \rho^2 - 2 \right) \zeta + 1 \right), \tag{11b}$$

$$F_1(\zeta) = \frac{1}{4} x_{b0} + \frac{b^2 e^{-h}}{4(1+b^2)} x_{b0}' + \frac{1}{4(1+b^2)} ((\zeta - 1) M(\zeta) + e^{-h}) u_{-1}. \tag{11c}$$

If we now take the projection of both sides of (11) on MH^2, we see that

$$X_2(\zeta) x_b' = \frac{b^2}{4(1+b^2)} (x_{b0}' - \zeta u_{-1}). \tag{12}$$

Thus from (11), (11a)-(11c) and (12), we have

$$X_1(\zeta) x_b = \frac{1}{4} x_{b0} + \frac{b^2(e^{-h} - M(\zeta))}{4(1+b^2)} x_{b0}' + \frac{((-1 + (1+b^2)\zeta) M(\zeta) + e^{-h})}{4(1+b^2)} u_{-1}. \tag{13}$$

It is easy to check that for $b^2/(1+b^2) \leqq \rho^2 \leqq 1$, $X_2(\zeta)$ has one root, r_2, inside the unit disc D, and that $X_1(\zeta)$ has both roots, r_1, r_1^{-1} on the unit circle ∂D. Therefore we have $x'_{b0} = r_2 u_{-1}$, and so

$$\exp\left(-2h\frac{r_1+1}{r_1-1}\right) = \frac{-1-b^2r_2+(1+b^2)r_1}{-1-b^2r_2+(1+b^2)r_1^{-1}}.$$

Hence from (13) we may conclude that

$$hy + \tan^{-1}\frac{y\sqrt{1+b^2}}{\sqrt{1-b^2y^2}} = \pi, \tag{14}$$

where $y := \sqrt{1/\rho^2 - 1}$, $0 \leqq y \leqq 1/b$, and $\tan^{-1} y\sqrt{1+b^2}/\sqrt{1-b^2y^2} \in [0, \pi/2]$. Note that for (14) to have a solution, we need $h \geqq \pi b/2$. Hence, if $h \leqq \pi b/2$ then $\mu = \|A\|_e = b/\sqrt{1+b^2}$; otherwise $\mu = \rho = 1/\sqrt{y^2+1}$ where y is the unique solution of (14) in the range $0 \leqq y \leqq 1/b$. Note that when $b = 0$, (14) becomes

$$hy + \tan^{-1} y = \pi, \qquad 0 \leqq y \leqq \infty,$$

which is exactly the same equation obtained previously in [9], [10], [16], and [21] for the 1-block problem. Clearly, as $b \uparrow \infty$, we have that $\mu \uparrow 1$. The physical meaning of this situation is that in this case we infinitely penalize the energy of the command signal u. Indeed, since P is already stable we are allowed to choose $C = 0$, which will make $u = 0$, and hence solve the problem. However, in this situation the tradeoff is that the energy of the worst error signal cannot be less than the energy of the disturbance signal d, so μ will be equal to one. Figure 3 gives an indication on how μ depends on the parameters b and h.

6. Concluding remarks. In this paper we have studied H^∞ optimization of multivariable distributed systems. We took the most general case of the standard H^∞ problem, namely, the so-called 4-block problem. Here, we developed a rank type formula for the computation of the eigenvalues of the operator A^*A. It is important to emphasize once more that the crucial steps of the procedure presented here are: (i) to do the factorizations (f1)-(f3), and (ii) to find $X_0^{(-1)}$. We refer to the paper [3] for the methods of performing these steps. From a computational point of view, the same method may

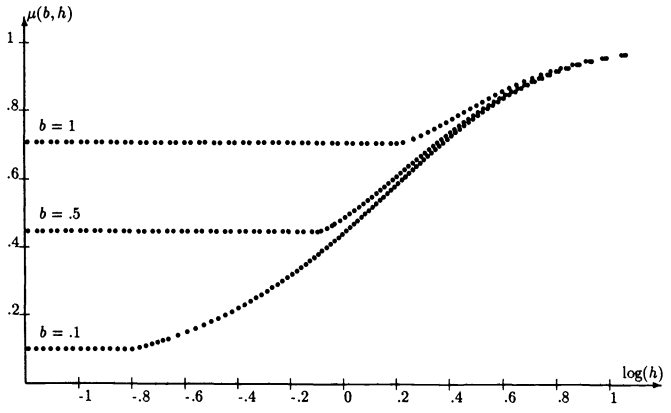


FIG. 3

be used to solve the 4-block problem for MIMO lumped systems, and MIMO stable distributed systems.

At this point we feel that the skew Toeplitz theory gives a satisfactory way of solving the optimal version of the 4-block problem in a very general setting. We should note that these techniques should also lead to the suboptimal solutions as considered in [2] and [4] for finite-dimensional systems using a state-space point of view. Indeed, since the operator A is derived from the commutant lifting theorem, we could in principle get all of the suboptimal solutions via the one-step extension technique of [1], once we know how to do the optimal case. This program has already been carried out for the 1-block case in [8]. Such a suboptimal parametrization would allow us to make contact with the very important work of [2] and [4]. Finally, it would be interesting to explore the possibility of combining state-space and frequency-domain methods in the 4-block problem as was done in [16] and [23] in the 1-block case.

REFERENCES

[1] V. M. ADAMJAN, D. Z. AROV, AND M. G. KREIN, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Shur-Takagi problem*, Math. USSR Sb., 15 (1971), pp. 31-73.

[2] J. A. BALL AND N. COHEN, *Sensitivity minimization in an H^∞ norm*, Internat. J. Control, 46 (1986), pp. 785-816.

[3] H. BERCOVICI, C. FOIAS, AND A. TANNENBAUM, *On skew Toeplitz operators*, Operator Theory: Adv. Appl., 32 (1988), pp. 21-43.

[4] J. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. FRANCIS, *State space solutions to standard H^2 and H^∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831-847.

[5] A. FEINTUCH AND B. FRANCIS, *Uniformly optimal control of linear systems*, Automatica, 21 (1986), pp. 563-574.

[6] C. FOIAS AND A. TANNENBAUM, *On the four block problem*, I, Operator Theory: Adv. Appl., 32 (1988), pp. 93-112.

[7] ———, *On the four block problem*, II: *the singular system*, Operator Theory and Integral Equations, 11 (1988), pp. 726-767.

[8] ———, *On the parametrization of the suboptimal solutions in generalized interpolation*, Linear Algebra Appl., 122/123/124 (1989), pp. 145-164.

[9] C. FOIAS, A. TANNENBAUM, AND G. ZAMES, *On the H^∞ optimal sensitivity problem for systems with delays*, SIAM J. Control Optim., 25 (1987), pp. 686-706.

[10] C. FOIAS AND A. TANNENBAUM, *On the Nehari problem for a certain class of L^∞ functions appearing in control theory*, J. Functional Anal., 74 (1987), pp. 146-159.

[11] C. FOIAS, A. TANNENBAUM, AND G. ZAMES, *Some explicit formulae for the singular values of certain Hankel operators with factorizable symbol*, SIAM J. Math. Anal., 19 (1988), pp. 1081-1091.

[12] C. FOIAS, H. ÖZBAY, AND A. TANNENBAUM, *Remarks on H^∞ optimization of multivariate distributed systems*, in Proc. Conference on Decision and Control, Austin, TX, 1988, pp. 985-986.

[13] B. FRANCIS, *A Course in H^∞ Control Theory*, Lecture Notes in Control and Information Sciences, 88, Springer-Verlag, Berlin, New York, 1987.

[14] E. JONCKHEERE AND M. VERMA, *A spectral characterization of H^∞ optimal feedback performance and its efficient computation*, Systems Control Lett., 8 (1985), pp. 13-22.

[15] P. P. KHARGONEKAR, H. ÖZBAY, AND A. TANNENBAUM, *A remark on the four block problem: stable plants and rational weights*, Internat. J. Control, 50 (1989), pp. 1013-1023.

[16] T. LYPCHUK, M. SMITH, AND A. TANNENBAUM, *Weighted sensitivity minimization: general plants in H^∞ and rational weights*, in Proc. Conference on Decision and Control, Los Angeles, CA, 1987; Linear Algebra Appl., 109 (1988), pp. 71-90.

[17] N. K. NIKOL'SKII, *Treatise on the Shift Operator*, Springer-Verlag, Berlin, New York, 1986.

[18] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

[19] M. VERMA AND E. JONCKHEERE, *L^∞ compensation with mixed sensitivity as a broadband matching problem*, Systems Control Lett., 4 (1984), pp. 125-129.

[20] M. VIDYASAGAR, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

[21] G. ZAMES, A. TANNENBAUM, AND C. FOIAS, *Optimal H^∞ interpolation: a new approach*, in Proc. Conference on Decision and Control, Athens, Greece, 1986, pp. 350-355.

[22] G. ZAMES AND S. K. MITTER, *A note on essential spectrum and norms of mixed Hankel-Toeplitz operators*, Systems Control Lett., 10 (1988), pp. 159-165.

[23] K. ZHOU AND P. P. KHARGONEKAR, *On the weighted sensitivity minimization problem for delay systems*, Systems Control Lett., 8 (1987), pp. 307-312.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 671-677, May 1990

011

AN EVASION GAME ON A FINITE TREE*

V. J. BASTON† AND F. A. BOSTOCK†

Abstract. The paper considers the following two-person zero-sum multistage game. An evader starts at a given vertex of a tree and, at discrete intervals of time, chooses either to move to one of the vertices adjacent to him or stay where he is. A gunner with a single bullet may, at each of the same discrete intervals of time, either fire the bullet at any of the vertices of the tree or hold his fire. The gunner always hits the vertex at which he aims and the bullet takes one unit of time to reach its target. The payoff to the gunner is 1 if he hits the evader, μ (where $|\mu|<1$) if he fires and misses, and 0 if he never fires. It is shown that, whatever vertex the evader starts at, the value of the game is $(1+\mu)/2$.

Key words. two-person game, zero-sum game, time lag system, recursive matrix game

AMS(MOS) subject classifications. 90D05, 90D20

1. Introduction. Firing games in which there is a time lag have attracted attention over a period of years. This class of games occurs in different guises in a variety of situations such as a bomber-battleship problem [5], [7], [8], and [9] or a tank maneuvering to avoid gunfire [10]. Other papers on this theme include [4], [13], and [14]. More recently there has been considerable interest in problems in which an evader, moving on a discrete set of points, tries to avoid being hit by a gunner. Lee [11], [12] has investigated the case where there is a safe point for the evader, while Baston and Bostock [1] have treated the case where the number of points is finite. In these papers the gunner has a given number of bullets to start with, but Bernhard, Colomb and Papavassilopoulos [2] consider the situation in which the gunner has an unlimited supply. In the above papers the points can all be thought of as distributed on a line. However, it was conjectured that the result in [1] could be extended to a finite tree and the purpose of this paper is to prove that conjecture. The techniques employed in this paper are broadly similar to those in [1], but the technology involved is fundamentally different.

Let $A_1, A_2, \cdots, A_n$ be the vertices of a finite tree. An evader starts at some given vertex A_s and, at discrete intervals of time $t=1, 2, \cdots$, chooses either to move to one of the vertices adjacent to him or to stay where he is. A gunner with a single bullet may, at each of the same discrete intervals of time, either fire the bullet at any one of the vertices A_r or hold his fire. It is assumed that the gunner always hits the vertex at which he aims and that the bullet takes one unit of time to reach its target. The payoff to the gunner is 1 if he hits the evader, μ (where $|\mu|<1$) if he fires and misses, and 0 if he never fires. The special case where the tree has only two terminal vertices was solved in [1]; note that we use terminal vertex for a vertex of valency one. As in this special case we will obtain a solution for our game by modelling it as a recursive matrix game. For the notation and a brief account of recursive matrix games, the reader is referred to [1]; in particular we will have occasion to use Lemma 1 in that paper. We will prove that no matter at which vertex the game starts the value is always $(1+\mu)/2$; we also give an optimal strategy for the evader and show how an ε-optimal strategy for the gunner may be determined.

Of course we do not need the theory of recursive games to see a strategy for the evader which holds the gunner's expectation down to $(1+\mu)/2$. He may, when at a

* Received by the editors November 21, 1988; accepted for publication (in revised form) August 11, 1989.
† Faculty of Mathematical Sciences, University of Southampton, Southampton SO9 5NH, United Kingdom.

vertex A_r at time t, simply choose two of his alternatives each with probability $\frac{1}{2}$. It is perhaps surprising that this bound is tight, since the evader could be starting at a vertex with high valency. The situation for the gunner is however far more complex. It turns out that our ε-optimal strategy for the gunner is actually independent of μ; we will only deal with the case $\mu = 0$ as the result can be extended to $|\mu| < 1$ exactly in the manner of [1]. Finally, in § 5 we point out that our methods can be used to solve the game in which the evader must move.

2. The model as a recursive game. Let Γ_r represent the game where the evader starts at the vertex A_r of the tree G with vertices $A_1, A_2, \cdots, A_n$. We assume throughout $n \geqq 3$ since when $n = 2$ the problem is trivial. If the evader is at a terminal vertex, then the gunner can ensure himself an expectation of $\frac{1}{2}$ by immediately firing at the terminal vertex or its adjacent vertex, each with probability $\frac{1}{2}$. Hence, the value of a game which starts at a terminal vertex is $\frac{1}{2}$; thus, we may regard our problem in terms of the recursive matrix game $\Gamma = (\Gamma_1, \Gamma_2, \cdots, \Gamma_n)$, where for each terminal vertex A_i $\Gamma_i = (\frac{1}{2})$ and for each nonterminal vertex A_r the matrix M_r for Γ_r is described as follows. The columns of M_r are indexed by the set Δ_r which consists of the vertices adjacent to A_r together with A_r itself; an element A_j of Δ_r represents the pure strategy for the evader in which he moves to the vertex A_j. The rows of M_r are indexed by Δ_r together with an additional symbol δ that represents the pure strategy for the gunner in which he does not fire. A row indexed by an element A_i of Δ_r represents the pure strategy for the gunner in which he fires at the vertex A_i. The entries of M_r are given by

$$M_r(A_i, A_j) = \begin{cases} 1 & \text{when } A_i = A_j \\ 0 & \text{when } A_i \neq A_j \end{cases}$$

and

$$M_r(\delta, A_j) = \begin{cases} \frac{1}{2} & \text{when } A_j \text{ is a terminal vertex} \\ \Gamma_j & \text{when } A_j \text{ is not a terminal vertex.} \end{cases}$$

At this juncture we find it convenient to have $A_1, \cdots, A_N$ as the nonterminal vertices of G and $A_{N+1}, \cdots, A_n$ as terminal vertices. For each real N-vector $W = (w_1, \cdots, w_N)$, let $M_r(W)$ denote the real matrix obtained by substituting w_i for each game component Γ_i which occurs in M_r. The value map V from real N-vectors to real N-vectors is defined by taking the rth component of $V(W)$ as the value of $M_r(W)$, regarded as an ordinary matrix game. Now let W_0 denote the zero N-vector and for $k = 1, 2, \cdots$, define $W_k = (w_1^k, \cdots, w_N^k)$ by $W_k = V(W_{k-1})$. Since all the number entries in all the matrices M_r are nonnegative, and because the value of an ordinary matrix game is an increasing function of its entries, it is easy to see by induction that the sequence W_k is increasing.

3. The value of the game. The next four lemmas enable us to show that each sequence w_r^k converges to $\frac{1}{2}$. The proof of Lemma 1 is routine and is left to the reader.

LEMMA 1. *Let the real matrix $A = (a_{ij})$, $i = 1, 2, \cdots, m+1, j = 1, 2, \cdots, m$ be given by,*

$$a_{ij} = \begin{cases} 1 & \text{when } i = j, \quad 1 \leqq i, j \leqq m \\ 0 & \text{when } i \neq j, \quad 1 \leqq i, j \leqq m \end{cases}$$

and

$$a_{m+1j} = a_j, \qquad 1 \leqq j \leqq m,$$

where $0 \leqq a_1 \leqq a_2 \leqq \cdots \leqq a_m$. Then as an ordinary matrix game the value v and optimal strategies are given as follows.

(i) *When* $\sum_{i=1}^{m} a_i \leqq 1$, $v = 1/m$ *and optimal strategies* $X^* = (x_1, \cdots, x_{m+1})$ *for the row player and* $Y^* = (y_1, \cdots, y_m)$ *for the column player are given by,*

$$x_i = 1/m \quad \text{for } i = 1, \cdots, m, \qquad x_{m+1} = 0$$

and

$$y_i = 1/m \quad \text{for } i = 1, \cdots, m.$$

(ii) *When* $\sum_{i=1}^{m} a_i \geqq 1$ *and* s *is defined by* $\sum_{i=1}^{s-1} a_i < 1 \leqq \sum_{i=1}^{s} a_i$ *then* $v = a_s/\rho$ *where* $\rho = 1 + (s-1)a_s - \sum_{i=1}^{s-1} a_i$, *and the optimal strategies* X^* *and* Y^* *are given by*

$$x_i = (a_2 - a_i)/\rho \quad \text{for } i = 1, \cdots, s-1$$

$$x_i = 0 \quad \text{for } i = s, s+1, \cdots, m \qquad x_{m+1} = 1/\rho$$

and

$$y_i = a_s/\rho \quad \text{for } i = 1, \cdots, s-1$$

$$y_s = \left\{1 - \sum_{i=1}^{s-1} a_i\right\} \Big/ \rho$$

$$y_i = 0 \quad \text{for } i = s+1, \cdots, m.$$

LEMMA 2. *Let* A_r *be any nonterminal vertex, then for all* k, $w_r^k < \frac{1}{2}$.

Proof. We will prove the result by induction on k, so for the inductive step suppose that for some k and for all nonterminal vertices A_i, $w_i^k < \frac{1}{2}$. Let A_r be any nonterminal vertex, then the number, m say, of vertices in Δ_r, is at least three. By the definition of W_k,

$$w_r^{k+1} = \text{val} \begin{bmatrix} 1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 1 \\ a_1 & a_2 & \cdot & \cdot & a_m \end{bmatrix},$$

where each a_j is either $\frac{1}{2}$ corresponding to a terminal vertex adjacent to A_r, or is w_r^k, or is some w_i^k corresponding to a nonterminal vertex adjacent to A_r. Without loss of generality, we may take $a_1 = w_r^k$. By our hypothesis $a_1 < \frac{1}{2}$ and $a_2, a_3 \leqq \frac{1}{2}$, so the evader can restrict the expected payoff to strictly less than $\frac{1}{2}$ by choosing columns 1, 2, and 3 with equal probability. Thus $w_r^{k+1} < \frac{1}{2}$, and the proof of the induction step is complete. Since for all nonterminal vertices A_i, $w_i^0 = 0 < \frac{1}{2}$, the result now follows by induction.

As we have noted earlier, the sequence W_k is increasing, so since it is bounded above (by Lemma 2), it converges, say to $(w_1, w_2, \cdots, w_N)$. We note here that for $r = 1, \cdots, N$, $0 < w_r \leqq \frac{1}{2}$. The upper bound follows from Lemma 2, and the lower bound holds since clearly $w_r^1 > 0$ from Lemma 1. We also observe that because the value of an ordinary matrix game is a continuous function of its entries, the vector $(w_1, w_2, \cdots, w_N)$ is a fixed point of the value map.

LEMMA 3. *Let* A_r *be a (nonterminal) vertex with* $w_r < \frac{1}{2}$ *and such that for all (nonterminal) vertices* A_s $w_r \leqq w_s$, *then* A_r *has at least two adjacent (nonterminal) vertices* A_i *and* A_j *say, for which* $w_i = w_j = w_r$.

Proof. Let A_r be a (nonterminal) vertex with the hypothesis of the lemma. By our note and observation immediately preceding Lemma 3, we have

$$(*) \qquad w_r = \text{val}\begin{bmatrix} 1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 1 \\ w_r & a_2 & \cdot & \cdot & a_m \end{bmatrix}, \qquad m \geqq 3$$

where each a_s is either $\frac{1}{2}$ corresponding to a terminal vertex adjacent to A_r or is some w_s corresponding to a nonterminal vertex adjacent to A_r, and the a_s's are ordered so that $0 < w_r \leqq a_2 \leqq a_3 \leqq \cdots \leqq a_m \leqq \frac{1}{2}$. In the ordinary matrix game, let X denote the mixed strategy for the row player whereby he plays each of rows 1 and 2 with a small probability $\varepsilon > 0$ (the exact value of which is indicated later) and row $m+1$ with probability $1-2\varepsilon$. If $E(X, s)$ denotes the expectation when X is used against column s, then

$$\begin{aligned} E(X,1) &= \varepsilon + (1-2\varepsilon)w_r = w_r + 2\varepsilon(\tfrac{1}{2} - w_r) > w_r, \\ E(X,2) &= \varepsilon + (1-2\varepsilon)a_2 = w_r + a_2 - w_r + 2\varepsilon(\tfrac{1}{2} - a_2) \\ &> w_r \text{ since } w_r < \tfrac{1}{2} \text{ and } w_r \leqq a_2 \leqq \tfrac{1}{2}, \end{aligned}$$

and for $3 \leqq s \leqq m$,

$$\begin{aligned} E(X,s) &= (1-2\varepsilon)a_n \geqq (1-2\varepsilon)a_3 \\ &= w_r + 2a_3((a_3 - w_r)/(2a_3) - \varepsilon) \\ &> w_r \text{ for a sufficiently small } \varepsilon, \text{ if } a_3 > w_r. \end{aligned}$$

Hence (*) can only be satisfied if $a_3 = w_r$ and the result follows.

LEMMA 4. *For each (nonterminal) vertex* A_r, $w_r = \frac{1}{2}$.

Proof. By Lemma 2, for each (nonterminal) vertex A_s, $w_s \leqq \frac{1}{2}$. Suppose that for the vertex A_r, $w_r < \frac{1}{2}$. Since the tree is finite we may choose A_r so that for all (nonterminal) vertices A_s, $w_r \leqq w_s$. Then by Lemma 3 there are two adjacent vertices A_i and A_j with $w_i = w_j = w_r$. Thus each vertex of the subgraph given by the vertices A_s with $w_s = w_r$ has valency at least 2 and so is not a tree [3, p. 8].

4. Gunner strategies. The remainder of the paper is concerned with the determination of an ε-optimal strategy for the gunner. This is essentially achieved as an application of our final lemma.

LEMMA 5. *Let* $\varepsilon > 0$ *satisfy* $3(\frac{1}{2} - \varepsilon) \geqq 1$, *and let* k *be large enough to ensure that for all (nonterminal) vertices* A_i, $w_i^k \geqq \frac{1}{2} - \varepsilon$. *For any nonterminal vertex* A_r *let* D *denote the set of suffixes of those* A_i *which are nonterminal vertices adjacent to* A_r; *then*

(i) *If there exists* i *in* D *with* $w_i^k \leqq w_r^k$ *then for all* j *in* D *not equal to* i, $w_r^k \leqq w_j^k$,

(ii) *If there exists* i *in* D *with* $w_i^k < w_r^k$ *then for all* j *in* D *not equal to* i, $w_r^k < w_j^k$.

Proof. Let ε and k satisfy the hypothesis of the lemma, and let A_r be any nonterminal vertex. As in the proof of Lemma 2,

$$w_r^{k+1} = \text{val}\begin{bmatrix} 1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 1 \\ a_1 & a_2 & \cdot & \cdot & a_m \end{bmatrix},$$

where we also have ordered the a_j's so that $0 \leqq a_1 \leqq a_2 \leqq \cdots \leqq a_m$, and because A_r is a nonterminal vertex $m \geqq 3$. By Lemma 2 $w_r^k < \frac{1}{2}$, so $a_1 \neq \frac{1}{2}$. Since for all (nonterminal)

vertices A_i $w_i^k \geqq \frac{1}{2}-\varepsilon$ and $3(\frac{1}{2}-\varepsilon)\geqq 1$, we have from Lemma 1

$$w_r^k \leqq w_r^{k+1} = a_3/\{1+2a_3-a_1-a_2\} \leqq a_3.$$

Now let i in D be such that $w_i^k < w_r^k$. If $2a_3-a_1-a_2>0$ then $w_r^k < a_3$, so $w_i^k = a_1$ and $w_r^k = a_2$. If $2a_3-a_1-a_2=0$ then $a_1=a_2=a_3$, so again $w_r^k = a_2$. Thus for all j in D not equal to i, $w_r^k \leqq w_j^k$.

Now let i in D be such that $w_i^k < w_r^k$; then certainly $2a_3-a_1-a_2>0$, so that $w_r^k(=a_2) < a_3$. Thus for all j in D not equal to i, $w_r^k < w_j^k$, and the proof of the lemma is complete.

For each j let w^j denote the minimum of the w_i^j over (nonterminal) vertices A_i. With the hypothesis of Lemma 5 concerning ε and k, let $A_{r_0}, A_{r_1}, \cdots, A_{r_p}, A_{r_{p+1}}$ be any path of distinct vertices with $w_{r_0}^k = w^k$ and $A_{r_{p+1}}$ a terminal vertex; then Lemma 5 shows that for each $i=0,1,\cdots,p-1$, $w_{r_i}^k \leqq w_{r_{i+1}}^k$ with equality up to some vertex A_{r_j} say, and strict inequality thereafter so that $w^k = w_{r_0}^k = \cdots = w_{r_j}^k < w_{r_{j+1}} < \cdots < w_{r_p}^k < \frac{1}{2}$. This means that we have a subtree G_0 where for each vertex A_r of G_0, $w_r^k = w^k$ and for each nonterminal vertex A_r of G in $G-G_0$, $w_r^k > w^k$. Assume that G_0 has at least three vertices, and let A_i be a terminal vertex of G_0. Let A_i, A_{i_1}, A_{i_2} be any path of three distinct vertices in G_0, and let $A_{j_1}, A_{j_2}, \cdots, A_{j_p}$ be the vertices in $G-G_0$ which are adjacent to A_i. Note that because no terminal vertex of G_0 is a terminal vertex of G the set of the A_{j_r}'s is not empty. Define α to be the minimum of the $w_{j_r}^k$'s or to be $\frac{1}{2}$ should all the A_{j_r}'s be terminal vertices of G. Now consider the vector $(w_1^{k+1}, w_2^{k+1}, \cdots, w_N^{k+1})$. By Lemma 1 and, since $w_i^k = w_{i_1}^k = w_{i_2}^k (= w^k)$,

$$\begin{aligned} w_i^{k+1} - w_{i_1}^{k+1} &= \alpha/\{1+2\alpha-2w^k\} - w^k \\ &= (\alpha - w^k)(1-2w^k)/\{1+2\alpha-w^k\} \\ &> 0 \quad \text{since } \alpha > w_i^k = w^k < \tfrac{1}{2}. \end{aligned}$$

Thus $w_i^{k+1} > w_{i_1}^{k+1}$. Notice in particular that for any path A_r, A_s, A_t of three distinct vertices of G_0, $w_s^{k+1} = w^k$. Let G_1 be the subtree of G_0, obtained by removing the terminal vertices of G_0. For each vertex A_r of G_1, $w_r^{k+1} = w^k$ and if $A_{s_0}, A_{s_1}, \cdots, A_{s_p}, A_{s_{p+1}}$ is any path of distinct vertices with $A_{s_{p+1}}$ a terminal vertex of G and A_{s_0} in G_1, then by Lemma 5 we have

$$w^k = w_{s_0}^{k+1} = \cdots = w_{s_j}^{k+1} < w_{s_{j+1}}^{k+1} < \cdots < w_{s_p}^{k+1} < \tfrac{1}{2},$$

where A_{s_j} is a terminal vertex of G_1. Since for all (nonterminal) vertices A_i of G, $w_i^{k+1} \geqq w_i^k \geqq \frac{1}{2}-\varepsilon$, we may again make use of Lemma 1 in passing to the vector $(w_1^{k+2}, w_2^{k+2}, \cdots, w_N^{k+2})$, and, provided G_1 has at least three vertices, we can remove its terminal vertices to obtain an analogous subtree G_2 of G_1. Continuing the process in the obvious manner we arrive, after a finite number of steps m, say, at a subtree G_m which has just either one or two vertices. Let $h=k+m$, then for each vertex A_r of G_m, $w_r^h = w^k (= w^h)$, and if $A_{t_0}, A_{t_1}, \cdots, A_{t_p}, A_{t_{p+1}}$ is any path of distinct vertices with $A_{t_{p+1}}$ a terminal vertex of G and A_{t_0} in G_m, then $w^k = w_{t_0}^h \leqq w_{t_1}^h < w_{t_2}^h < \cdots < w_{t_p}^h < \frac{1}{2}$, and the first inequality is also strict unless A_{t_1} is the possible second vertex in G_m.

We are now in a position to construct a stationary strategy $X(\varepsilon)$ for the gunner which will ensure that his expectation is at least $\frac{1}{2}-\varepsilon$ irrespective of what the evader does. In the ordinary matrix game $M_r(W_h)$, let X_r be the optimal strategy for the row player which is given by applying Lemma 1, and define the stationary strategy $^*X$ for Player 1 in the recursive game by taking $^*X_r^t = X_r$ for all t. Since $V(W_h) \geqq W_h$, the condition (i) of Lemma 1 in [1] regarding $^*X$ and W_h is satisfied. By the work following Lemma 5 regarding paths emanating from a vertex in G_m, we may conclude that for the nonterminal vertex A_r there exist adjacent (nonterminal) vertices A_p and A_q such

that either, (i) $w_p^h \leqq w_r^h < w_q^h$, and for all other adjacent vertices A_s, $w_q^h \leqq w_s^h$ or (ii) $w_p^h \leqq w_r^h$, and all other adjacent vertices are terminal. Thus, in the optimal strategy X_r, the probability of not firing (namely, $1/\{1+2w_q^h - w_p^h - w_r^h\}$ in the event of (i) and $1/\{1+2(\frac{1}{2}) - w_p^h - w_r^h\}$ in the event of (ii)) is strictly less than one. This is true at each nonterminal vertex, so since the tree is finite there exists $\alpha < 1$ such that at all nonterminal vertices the probability of not firing in an optimal strategy is never greater than α. This means that for all strategies Y for Player 2 in the recursive game and for all t, each matrix $Q^t(*X, Y)$ (see Lemma 1 in [1]) has each of its row sums at most α. It is an easy exercise to see that condition (ii) of Lemma 1 in [1] is also satisfied. Hence the strategy $*X$ will ensure the gunner has an expectation of at least $\frac{1}{2} - \varepsilon$, and our investigation for the case $\mu = 0$ is completed.

5. Conclusions. It is perhaps of interest to point out some of the properties of the gunner strategy $*X$ we obtained in the previous section. When ε is small the gunner fires with only a small probability when he sees the evader at a nonterminal vertex. Furthermore, even if this nonterminal vertex has a very high valency, our analysis shows that there is at most one adjacent vertex where the evader would be at least as well off as his present position; at all other adjacent vertices the evader would be worse off. Not surprisingly, therefore, our strategy tells the gunner to shoot with nonzero probability only at the vertex where the evader is and at most one of the adjacent vertices. The extension in [1] to the gunner having j bullets also applies here.

Our analysis may easily be adapted to show that, in the game where the evader must move, the gunner may hit the evader with probability as close to one as desired, no matter where the evader starts. In particular the adjustment of the gunner strategy $*X$ causes no difficulty. It is natural to consider generalisations to arbitrary graphs. For such games each vertex gives rise, in an obvious manner, to a component matrix of a recursive game. An iteration of the zero vector under the value map will again converge to the value of the game. Furthermore, as in [6] an ε-optimal or optimal strategy for the gunner can be derived from the iterative process, but the strategy so obtained is not necessarily stationary. Also, although this process may solve an

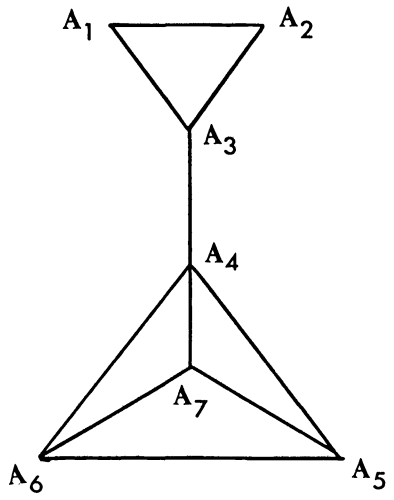


FIG. 1

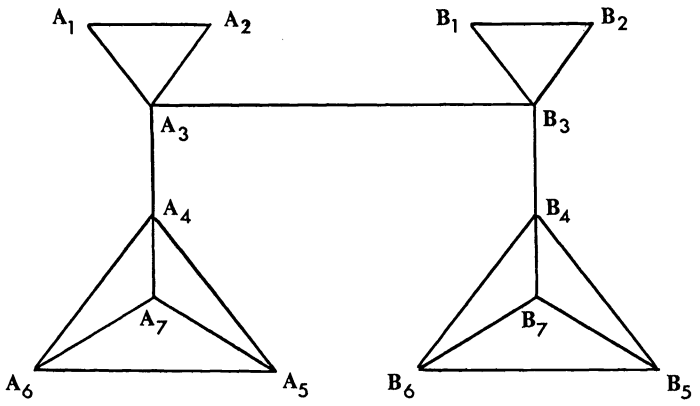


FIG. 2

individual case, it would not, per se, show the general influence of graph structure on the value. For instance, when the graph is not a tree, the components of the value may not all be equal. In illustrating some of these points we shall restrict our attention to games where the evader must move.

For the graph in Fig. 1 it is easy to show that the value v is given by $v_1 = v_2 = \frac{1}{2}$, $v_3 = \frac{3}{7}$, and $v_i = \frac{1}{3}$ otherwise. Now consider the graph in Fig. 2 where it is considerably more difficult to establish that the value $v = (v^A, v^B)$ is given by $v_1^A = v_2^A = \frac{1}{2}$, $v_3^A = (5-\sqrt{7})/6$, $v_i^A = \frac{1}{3}$ otherwise, and $v_i^A = v_i^B$ for all i. It appears from these two examples that a comprehensive structure theory relating the value of a graph to the value of its "constituent parts" in unlikely.

REFERENCES

[1] V. J. BASTON AND F. A. BOSTOCK, *An evasion game with barriers*, SIAM J. Control Optim., 26 (1988), pp. 1099-1105.

[2] P. BERNHARD, A.-L. COLOMB, AND G. P. PAPAVASSILOPOULOS, *Rabbit and hunter game: two discrete stochastic formulations*, Comput. Math. Appl., 13 (1987), pp. 205-225.

[3] B. BOLLOBÁS, *Graph Theory. An Introductory Course*, Graduate Texts in Mathematics 63, Springer-Verlag, New York, 1979.

[4] J. L. BURROW, *A multistage game with incomplete information requiring an infinite memory*, J. Optim. Theory Appl., 24 (1978), pp. 337-360.

[5] L. E. DUBINS, *A discrete evasion game*, in Ann. Math. Stud. 39, Princeton University Press, Princeton, NJ, 1957, pp. 231-255.

[6] H. EVERETT, *Recursive games*, in Ann. Math. Stud. 39, Princeton University Press, Princeton, NJ, 1957, pp. 47-78.

[7] T. FERGUSON, *On discrete evasion games with a two-move information lag*, in Proc. 5th Berkeley Symp. Math. Stat. Probability Vol. 1, 1967, pp. 453-462.

[8] R. ISAACS, *The problem of aiming and evasion*, Naval Research Logistics Quarterly, 2 (1955), pp. 47-67.

[9] S. KARLIN, *An infinite move game with a lag*, in Ann. Math. Stud. 39, Princeton University Press, Princeton, NJ, 1957, pp. 257-272.

[10] P. R. KUMAR, *Optimal mixed strategies in a dynamic game*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 743-749.

[11] K. T. LEE, *A firing game with time lag*, J. Optim. Theory Appl., 41 (1983), pp. 547-558.

[12] ———, *An evasion game with a destination*, J. Optim. Theory Appl., 46 (1985), pp. 359-372.

[13] D. MATULA, *1100 and other embedded sequence games*, Mimeographed Notes, Berkeley, CA, 1964.

[14] H. E. SCARF AND L. S. SHAPLEY, *Games with partial information*, in Ann. Math. Stud. 39, Princeton University Press, Princeton, NJ, 1957, pp. 213-229.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 678-710, May 1990

012

PARTIALLY ASYNCHRONOUS, PARALLEL ALGORITHMS FOR NETWORK FLOW AND OTHER PROBLEMS*

P. TSENG†, D. P. BERTSEKAS†, AND J. N. TSITSIKLIS†

Abstract. The problem of computing a fixed point of a nonexpansive function f is considered. Sufficient conditions are provided under which a parallel, partially asynchronous implementation of the iteration $x := f(x)$ converges. These results are then applied to (i) quadratic programming subject to box constraints, (ii) strictly convex cost network flow optimization, (iii) an agreement and a Markov chain problem, (iv) neural network optimization, and (v) finding the least element of a polyhedral set determined by a weakly diagonally dominant, Leontief system. Finally, simulation results illustrating the attainable speedup and the effects of asynchronism are presented.

Key words. parallel algorithms, asynchronous algorithms, nonexpansive functions, network flows, neural networks, agreement, Markov chains, Leontief systems

AMS(MOS) subject classifications. 49, 90

1. Introduction.

In this paper we consider the computation of a fixed point of a nonexpansive function f using parallel, partially asynchronous iterative algorithms of the form $x := f(x)$. We give sufficient conditions under which such algorithms converge, we show that some known methods satisfy these conditions, and we propose some new algorithms. The convergence behavior of our methods is qualitatively different from the convergence behavior of most asynchronous algorithms that have been studied in the past by many authors [1]-[3], [5], [8], [27]-[30].

We consider a fixed point problem in the n-dimensional Euclidean space $\Re^n$. We are given functions $f_i : \Re^n \to \Re$, $i = 1, \cdots, n$, and we wish to find a point $x^* \in \Re^n$ such that

$$x^* = f(x^*),$$

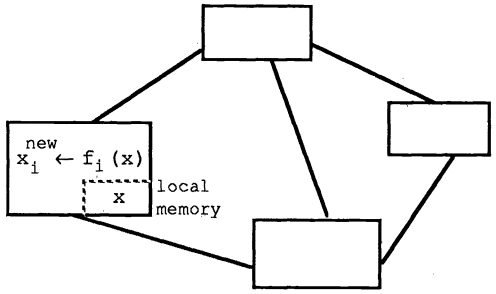
where $f : \Re^n \to \Re^n$ is defined by $f(x) = (f_1(x), \cdots, f_n(x))$.

We consider a network of processors endowed with local memories, which communicate by message passing, and which do not have access to a global clock. We assume that there are exactly n processors, each of which maintains its own estimate of a fixed point, and that the ith processor is responsible for updating x_i, the ith component of x. (If the number of processors is smaller than n, we may let each processor update several components; the mathematical description of the algorithm does not change and our results apply to this case as well.) We assume that processor i updates its component by occasionally applying f_i to its current estimate, say x, and then transmitting (possibly with some delay) the computed value $f_i(x)$ to all other processors, which use this value to update the ith component of their own estimates (see Fig. 1.1).

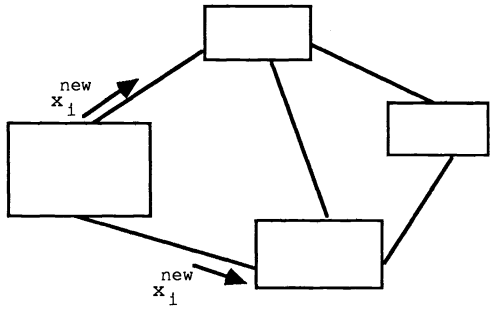
We use a nonnegative integer variable t to index the events of interest (e.g., processor updates). We will refer to t as *time*, although t need not correspond to the time of a global clock. We use the following notations:

* Received by the editors November 14, 1988; accepted for publication (in revised form) July 21, 1989.

† Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This work was supported by National Science Foundation grants NSF-ECS-8519058 and NSF-ECS-8552419, with matching funds from Bellcore and Du Pont, and by Army Research Office grant DAAL03-86-K-0171.



(a)



(b)

FIG. 1.1. (a) *Processor i computes new estimate of the* ith *component of a fixed point.* (b) *Processor i transmits new estimate to other processors.*

$x_i(t) =$ ith component of the solution estimate stored by processor i at time t.

$\mathcal{T}_i =$ an infinite set of times at which processor i updates x_i.

$\tau_{ij}(t) =$ a time at which the jth component of the solution estimate stored by processor i at time t was stored in the local memory of processor j ($j = 1, \cdots, n$; $t \in \mathcal{T}_i$). (Naturally, $\tau_{ij}(t) \leqq t$.)

In accordance with the above definitions, we postulate that the variables $x_i(t)$ evolve according to:

$$x_i(t+1) = \begin{cases} f_i(x_1(\tau_{i1}(t)), \cdots, x_n(\tau_{in}(t))) & \text{if } t \in \mathcal{T}_i, \\ x_i(t) & \text{otherwise.} \end{cases} \tag{1.1}$$

The initial conditions $x_i(0)$ are given, and for notational convenience we assume that $x_i(t) = x_i(0)$ for $t \leqq 0$, so that the asynchronous iteration (1.1) is well defined for $\tau_{ij}(t) \leqq 0$. We may view the difference $t - \tau_{ij}(t)$ as a "communication delay" between the current time t and the time $\tau_{ij}(t)$ at which the value of the jth coordinate, used by processor i at time t, was generated at processor j.

Asynchronous computation models may be divided into *totally asynchronous* and *partially asynchronous.* In the totally asynchronous model [1]-[3], [8], [30], the "delays" $t - \tau_{ij}(t)$ can become unbounded as t increases. This is the main difference with the partially asynchronous model, where the amounts $t - \tau_{ij}(t)$ are assumed bounded; in particular, the following assumption holds.

Assumption A. (Partial Asynchronism). There exists a positive integer B such that, for each i and each $t \in \mathcal{T}_i$, there holds:

(a) $0 \leqq t - \tau_{ij}(t) \leqq B - 1$, for all $j \in \{1, \cdots, n\}$.

(b) There exists $t' \in \mathcal{T}_i$ for which $1 \leqq t' - t \leqq B$.

(c) $\tau_{ii}(t) = t$.

Parts (a) and (b) of Assumption A state that both the communication delays and the processor idle periods are bounded and can be expected to hold in most practical cases; for example, (b) holds if each processor uses a local clock, if the ratio of the speeds of different local clocks is bounded, and if each processor computes periodically according to its own local clock (see [7], p. 484). Part (c) of Assumption A states that a processor i always uses the most recent value of its own component x_i. This assumption typically holds in practice, but it is interesting to note that, while it is necessary for our results (see the proof of Lemma 2.3(a)), it is not needed in the convergence analysis of totally asynchronous algorithms.

Partially asynchronous iterations have already been studied in the context of gradient optimization algorithms, for which it was shown that convergence is obtained provided that the bound B of Assumption A is sufficiently small [27]–[29]. Our results concern a fundamentally different class of partially asynchronous methods which are convergent for *every* value of the bound B. At least two interesting examples of such methods are known: the agreement algorithm of [29] and the Markov chain algorithm of [20]. However, it appears that these methods have not been recognized earlier as a class. Their convergence behavior is somewhat surprising because their totally asynchronous versions do not converge in general; for a counterexample, see [7, p. 484].

In this paper we focus on the convergence issues of partially asynchronous methods with arbitrarily large values of the asynchronism bound B. Our main result (Proposition 2.1) is the first general convergence result for these methods. In §§ 3–7, we show that Proposition 2.1 applies to a variety of methods for several important problems, including the agreement and Markov chain algorithms mentioned earlier. Some of our convergence results are new, even when they are specialized to the case of synchronous algorithms; for example, the convergence of Jacobi relaxation methods for strictly convex cost network flow problems in § 4.

2. A general convergence theorem. Throughout this paper, we let $X^* = \{x \in \Re^n \mid f(x) = x\}$ be the set of fixed points of f and, for each $x \in \Re^n$, we let $\|x\| = \max_{i=1,\cdots,n} |x_i|$ denote the maximum norm of x. For any $x \in \Re^n$, we denote by $\rho(x)$ the distance of x from X^*, defined by

$$\rho(x) = \inf_{y \in X^*} \|x - y\|.$$

Finally, given any $x \in \Re^n$ and $x^* \in X^*$, we let $I(x; x^*)$ be the set of indices of coordinates of x that are farthest away from x^*, that is,

$$I(x; x^*) = \{i \mid |x_i - x_i^*| = \|x - x^*\|\},$$

and we also denote

$$U(x; x^*) = \{y \in \Re^n \mid y_i = x_i \text{ for all } i \in I(x; x^*),$$
$$\text{and } |y_i - x_i^*| < \|x - x^*\| \text{ for all } i \notin I(x; x^*)\}.$$

Loosely speaking, $U(x; x^*)$ is the set of all vectors y with $\|y - x^*\| = \|x - x^*\|$ that agree with x in the components that are farthest away from x^* (see Fig. 2.1).

Our main assumption on the structure of f is the following.

Assumption B.

(a) f is continuous.

(b) The set of fixed points X^* is convex and nonempty.

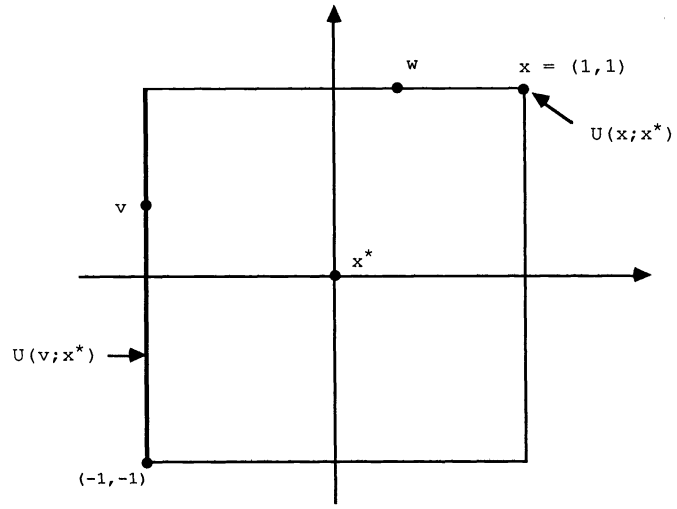


FIG. 2.1. *Illustration of the sets* $I(\cdot; x^*)$ *and* $U(\cdot; x^*)$. *Let* $n=2$ *and suppose that* $x^*=(0,0)\in X^*$. *For the indicated points* x, v, *and* w, *we have* $I(x; x^*)=\{1,2\}$, $I(v; x^*)=\{1\}$, $I(w; x^*)=\{2\}$. *The set* $U(v; x^*)$ *is the set of all vectors of the form* $(-1, c)$, *where* c *satisfies* $-1<c<1$, *which is the segment joining the points* $(-1,-1)$ *and* $(-1,1)$, *the endpoints excluded. Similarly,* $U(w; x^*)=\{(c,1)\mid -1<c<1\}$. *Finally, we have* $U(x; x^*)=\{x\}$.

(c) $\|f(x)-x^*\|\leqq\|x-x^*\|$, for all $x\in\Re^n$, for all $x^*\in X^*$.

(d) For every $x\in\Re^n$ and $x^*\in X^*$ such that $\|x-x^*\|=\rho(x)>0$, there exists some $i\in I(x; x^*)$ such that $f_i(y)\neq y_i$ for all $y\in U(x; x^*)$.

Part (c) of Assumption B states that f does not increase the distance from a fixed point and will be referred to as the *pseudo-nonexpansive* property. This is slightly weaker than requiring that f be nonexpansive (that is, $\|f(x)-f(y)\|\leqq\|x-y\|$ for all x and y in $\Re^n$) and in certain cases is easier to verify (see § 4). We interpret part (d) as follows: Consider some $x\notin X^*$. Then $f(x)\neq x$, and there exists some i such that $f_i(x)\neq x_i$. Assumption B(d) imposes the additional requirement that such an i can be found among the set of worst indices, that is, i belongs to the set $I(x; x^*)$ of indices corresponding to components farthest away from a closest element of X^*. Furthermore, if we change some of the other components of x to obtain another vector $y\in U(x; x^*)$, we still retain the property $f_i(y)\neq y_i$, for this particular i. This part of Assumption B is usually the most difficult to verify in specific applications.

Unfortunately, the following simple example shows that Assumptions A and B alone are not sufficient for convergence of even the synchronous version of iteration (1.1): Suppose that $f(x_1, x_2)=(x_2, x_1)$ (which can be verified to satisfy Assumption B with $X^*=\{(\lambda,\lambda)\mid\lambda\in\Re\}$). Then the sequence $\{x(t)\}$ generated by the synchronous iteration $x(t+1)=f(x(t))$ (which is a special case of (1.1)), with $x(0)=(1,0)$, oscillates between $(1,0)$ and $(0,1)$.

The difficulty in this example is that, at each iteration, while the worst coordinate $i\in I(x; x^*)$ is changed from 1 to 0, the other coordinate is increased from 0 to 1, and the distance $\rho(x)$ from X^* is not changed. The following assumption is designed to prevent such behavior.

Assumption C. For any i, $x\in\Re^n$, and $x^*\in X^*$, if $f_i(x)\neq x_i$, then $|f_i(x)-x_i^*|<\|x-x^*\|$.

An important fact, shown below, is that any mapping satisfying Assumption B can be modified by introducing a relaxation parameter, so that it satisfies Assumption C as well.

LEMMA 2.1. *Let $h: \Re^n \to \Re^n$ be a function satisfying Assumption* B. *Then the mapping $f: \Re^n \to \Re^n$ whose* ith *component is*

$$f_i(x) = (1-\gamma_i)x_i + \gamma_i h_i(x),$$

where $\gamma_1, \cdots, \gamma_n$ are scalars in $(0, 1)$, has the same set of fixed points as h and satisfies both Assumptions B *and* C.

Proof. It is easily seen that f is continuous and has the same set of fixed points as h, so it satisfies parts (a) and (b) of Assumption B. Since $f_i(x) \neq x_i$ if and only if $h_i(x) \neq x_i$, we see that f satisfies part (d) of Assumption B. Since h is pseudo-nonexpansive, for all i, $x \in \Re^n$, and $x^* \in X^*$, both x_i and $h_i(x)$ belong to the interval

$$[x_i^* - \|x - x^*\|, x_i^* + \|x - x^*\|].$$

Therefore, $f_i(x)$, which is a convex combination of x_i and $h_i(x)$, must also belong to this interval, proving that f is pseudo-nonexpansive, (cf. part (c) of Assumption B). Furthermore, if $h_i(x) \neq x_i$, then the convex combination $f_i(x)$ must belong to the interior of this interval, showing that f satisfies Assumption C. □

We now prove our main convergence result, showing that Assumptions A, B, and C are sufficient for the sequence $\{x(t)\}$ generated by the asynchronous iteration (1.1) to converge to an element of X^*. To motivate our proof, consider the synchronous iteration $x(t+1) = f(x(t))$. Under Assumptions B and C, either (i) $\rho(x(t+1)) < \rho(x(t))$ or (ii) $\rho(x(t+1)) = \rho(x(t))$ and $x(t+1)$ has a smaller number of components at a distance $\rho(x(t))$ from X^* than $x(t)$. Thus, case (ii) can occur for at most n successive iterations before case (i) occurs. This argument can be extended for the asynchronous iteration (1.1), but because of communication and computation delays (each bounded by B, due to Assumption A), the number of time steps until the distance to X^* decreases is upper bounded by roughly $2nB$ (see part (c) of Lemma 2.3).

PROPOSITION 2.1. *Suppose that $f: \Re^n \to \Re^n$ satisfies Assumptions* B *and* C, *and suppose that Assumption* A (*partial asynchronism*) *holds. Then the sequence $\{x(t)\}$ generated by the asynchronous iteration* (1.1) *converges to some element of X^*.*

Proof. For each integer $t \geqq 0$ denote

$$z(t) = (x(t-B+1), \cdots, x(t)),$$

$$d(z(t)) = \min_{x^* \in X^*} \{\max \{\|x(t-B+1) - x^*\|, \cdots, \|x(t) - x^*\|\}\}.$$

Notice that the minimum in the definition of $d(z(t))$ is attained because the set X^* is closed (as a consequence of the continuity of f). For each $t \geqq 0$, we fix an element $x^*(t)$ of X^* attaining the minimum

$$x^*(t) = \arg\min_{x^* \in X^*} \{\max \{\|x(t-B+1) - x^*\|, \cdots, \|x(t) - x^*\|\}\}. \tag{2.1}$$

As part of the proof of Proposition 2.1, we prove some preliminary facts in the following two lemmas, which show that the distance $d(z(t))$ cannot increase at any iteration while it decreases strictly "every few" iterations.

LEMMA 2.2. $d(z(t+1)) \leqq d(z(t))$, *for all* $z(t) \in \Re^{nB}$, *for all* $t \geqq 0$.

Proof. We will prove by induction that

$$\|x(r) - x^*(t)\| \leqq d(z(t)), \quad \forall r \geqq t-B+1, \tag{2.2}$$

which implies the result. From (2.1) and the definition of $d(z(t))$, this inequality holds for $r \in \{t-B+1, \cdots, t\}$. Suppose that it holds for all $r \in \{t-B+1, \cdots, r'\}$, where r' is some integer greater than or equal to t. We will show that it holds for $r'+1$. By (1.1), for each i, either $x_i(r'+1) = x_i(r')$ or $x_i(r') = f_i(x_1(\tau_{i1}(r')), \cdots, x_n(\tau_{in}(r')))$. In the former case, we have $|x_i(r'+1) - x_i^*(t)| = |x_i(r') - x_i^*(t)| \leqq d(z(t))$ by the induction hypothesis. In the latter case, we have by Assumption A(a), $r'-B+1 \leqq \tau_{ij}(r') \leqq r'$, so by the induction hypothesis, $|x_j(\tau_{ij}(r')) - x_j^*(t)| \leqq d(z(t))$ for all j. Using the pseudo-nonexpansive property of Assumption B(c), we obtain

$$|x_i(r'+1) - x_i^*(t)| \leqq \max_j |x_j(\tau_{ij}(r')) - x_j^*(t)| \leqq d(z(t)).$$

Thus, in either case we have $|x_i(r'+1) - x_i^*(t)| \leqq d(z(t))$, and this is true for every index i. Therefore, $\|x(r'+1) - x^*(t)\| \leqq d(z(t))$, completing the induction. □

LEMMA 2.3. *Fix some $t \geqq 0$ for which $d(z(t)) > 0$ and denote*

$$J(r) = \{i \,|\, |x_i(r) - x_i^*(t)| = d(z(t))\}, \quad \forall r \geqq t.$$

(a) *If $x_i(r+1) \neq x_i(r)$ for some $r \geqq t$, then $i \notin J(r+1)$.*

(2.3) (b) *$J(r+1) \subseteq J(r)$, for all $r \geqq t$.*

(c) $d(z(t+2nB+B-1)) < d(z(t))$.

Proof. For convenience, we will use the notation

$$\beta = d(z(t)), \qquad x^* = x^*(t).$$

(a) If $x_i(r+1) \neq x_i(r)$, we have $r \in \mathcal{T}_i$. Furthermore,

$$f_i(x_1(\tau_{i1}(r)), \cdots, x_n(\tau_{in}(r))) = x_i(r+1) \neq x_i(r) = x_i(\tau_{ii}(r)),$$

where the last equality follows from Assumption A(c). Using Assumption C, we obtain

$$|x_i(r+1) - x_i^*| < \max_j |x_j(\tau_{ij}(r)) - x_j^*| \leqq \beta,$$

where the last inequality follows from $r-B+1 \leqq \tau_{ij}(r) \leqq r$ (cf. Assumption A(a)) and Lemma 2.2 (cf., (2.2)). Thus, $i \notin J(r+1)$.

(b) If $i \in J(r+1)$, then part (a) shows that $x_i(r) = x_i(r+1)$, which implies that $i \in J(r)$.

(c) We first show by contradiction that, for all $r \geqq t$,

$$d(z(r+2B)) = \beta \Rightarrow J(r+2B) \neq J(r). \tag{2.4}$$

Suppose that, for some $r \geqq t$, we have $d(z(r+2B)) = \beta$ and $J(r) = J(r+2B)$. By part (b), $J(r) = J(r+1) = \cdots = J(r+2B)$. Denote $J = J(r)$. Then, by part (a),

$$x_i(r) = x_i(r+1) = \cdots = x_i(r+2B), \quad \forall i \in J, \tag{2.5}$$

and by the definition of J,

$$|x_i(r) - x_i^*| < \beta, \cdots, |x_i(r+2B) - x_i^*| < \beta, \quad \forall i \notin J. \tag{2.6}$$

Now, from the definition of J, x^* and β we have that $|x_i(r) - x_i^*| = \beta$ for all $i \in J$; hence (2.6) implies

$$\|x(r) - x^*\| = \beta, \qquad J = I(x(r); x^*). \tag{2.7}$$

Also by Assumption A(b), for each $i \in J$, there exists $r_i \in \{r+B, \cdots, r+2B-1\}$ such that $r_i \in \mathcal{T}_i$ and the iteration (1.1) yields

$$x_i(r_i+1) = f_i(x_1(\tau_{i1}(r_i)), \cdots, x_n(\tau_{in}(r_i))), \quad \forall i \in J. \tag{2.8}$$

Let us denote

$$x^i = (x_1(\tau_{i1}(r_i)), \cdots, x_n(\tau_{in}(r_i))), \quad \forall i \in J.$$

By Assumption A(c), $\tau_{ii}(r_i) = r_i$ for all $i \in J$, which together with (2.5) implies that

$$x_i(r_i+1) = x_i(\tau_{ii}(r_i)), \quad \forall i \in J.$$

Therefore, (2.8) can be written as

$$x_i^i = f_i(x^i), \quad \forall i \in J. \tag{2.9}$$

Furthermore, by Assumption A(a), $r \leqq \tau_{ij}(r_i) \leqq r+2B$ for all $i \in J$ and all j, which together with (2.5)-(2.6) implies that

$$x_j^i = x_j(r), \qquad \forall i \in J, \quad \forall j \in J,$$
$$|x_j^i - x_j^*| < \beta, \quad \forall i \in J, \quad \forall j \notin J.$$

Therefore from (2.7) we also have

$$x^i \in U(x(r); x^*), \quad \forall i \in J. \tag{2.10}$$

It now follows that

$$\beta = \|x(r) - x^*\| > \rho(x(r)),$$

since if $\|x(r) - x^*\| = \rho(x(r))$, then in view of the fact $I(x(r); x^*) = J$ (cf. (2.7)) and (2.9)-(2.10), Assumption B(d) would be violated.

Thus, we conclude that there exist $y^* \in X^*$ and $\theta \in [0, \beta)$ such that $\|x(r) - y^*\| = \theta$. Let

$$\varepsilon = \max\{|x_i(m) - x_i^*| \mid i \notin J,\ m = r+B, \cdots, r+2B-1\},$$
$$M = \max\{|x_i(m) - y_i^*| \mid i \notin J,\ m = r+B, \cdots, r+2B-1\}$$

(see Fig. 2.2). Since X^* is convex, we have that, for any $\omega \in (0, 1)$, $z^* = (1-\omega)x^* + \omega y^*$

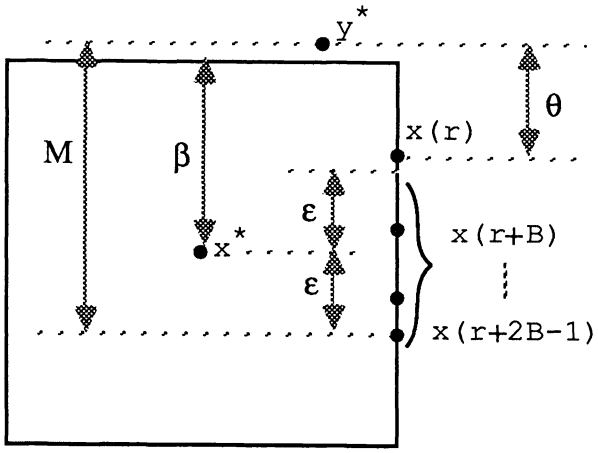


FIG. 2.2

is in X^* and, for $m = r+B, \cdots, r+2B-1$,

$$\begin{aligned}|x_i(m)-z_i^*| &= |x_i(r)-z_i^*| \\ &\leqq (1-\omega)|x_i(r)-x_i^*|+\omega|x_i(r)-y_i^*| \\ &= (1-\omega)\beta+\omega\theta, \quad \forall i \in J, \\ |x_i(m)-z_i^*| &\leqq (1-\omega)|x_i(m)-x_i^*|+\omega|x_i(m)-y_i^*| \\ &\leqq (1-\omega)\varepsilon+\omega M, \quad \forall i \notin J.\end{aligned}$$

Since $\varepsilon < \beta$ and $\theta < \beta$, we have that, for ω sufficiently small,

$$\|x(m)-z^*\| < \beta, \quad \forall m = r+B, \cdots, r+2B-1.$$

This implies that $d(z(r+2B-1)) < \beta$, a contradiction.

Since by Lemma 2.2, $d(z(r))$ is nonincreasing, either $d(z(t+2nB-1)) < \beta$, in which case the result is proved, or $d(z(t+2nB-1)) = \beta$. In the latter case, by (2.3) and (2.4), $J(t+2nB) = \cdots = J(t+2nB+B-1) = \varnothing$, and

$$d(z(t+2nB+B-1)) = \max\{\|x(t+2nB)-x^*\|, \cdots, \|x(t+2nB+B-1)-x^*\|\} < \beta.$$

□

We now complete the proof of Proposition 2.1.

By (2.2), the sequence $\{z(t)\}$ is bounded and, by Lemma (2.3)(c), $d(z(t))$ monotonically decreases to some limit β. If $\beta = 0$, then Lemma 2.2 and (2.2) imply that $\{x(r)\}$ has a unique limit point, which is in X^*, and our proof is complete. Suppose, to obtain a contradiction, that $\beta > 0$. Let

$$\Delta t = 2nB+B-1.$$

Since, by (2.2), $\{z(t)\}$ is bounded, there exist some $z^* \in \Re^{nB}$, $z^{**} \in \Re^{nB}$ and a subsequence T of $\{0, 1, \cdots\}$ such that

$$\{z(t)\}_{t\in T} \to z^*, \qquad \{z(t+\Delta t)\}_{t\in T} \to z^{**}. \tag{2.11}$$

Note that since $d(z(t)) \to \beta$ and d is a continuous function, (2.11) implies that $d(z^*) = d(z^{**}) = \beta$.

From (1.1), Assumption A and the definition of $z(t)$, we see that we can express $z(t+\Delta t)$ as a continuous function of $z(t)$. In particular, we can write

$$z(t+\Delta t) = g(z(t); \Gamma(t)), \tag{2.12}$$

where $\Gamma(t) = (\Gamma_1(t), \cdots, \Gamma_n(t))$ and $\Gamma_i(t)$ denotes the set

$$\Gamma_i(t) = \{(r-t, \tau_{i1}(r)-t, \cdots, \tau_{in}(r)-t) \mid r \in \mathcal{T}_i \cap \{t, \cdots, t+\Delta t\}\}, \tag{2.13}$$

and $g(\cdot; \Gamma(t)) : \Re^{nB} \to \Re^{nB}$ is some continuous function that depends on f and $\Gamma(t)$ only. (Note that $g(\cdot; \Gamma(t))$ is the composition of the f_i's in an order determined by $\Gamma(t)$ and is continuous because f is continuous.) Since (cf. (2.13) and Assumption A) $\Gamma(t)$ takes values from a *finite* set, by further passing into a subsequence, if necessary, we can assume that $\Gamma(t)$ is the same set for all $t \in T$. Let $\Gamma = (\Gamma_1, \cdots, \Gamma_n)$ denote this set. Then from (2.12) we obtain that

$$z(t+\Delta t) = g(z(t); \Gamma), \quad \forall t \in T.$$

Since $g(\cdot; \Gamma)$ is continuous, this, together with (2.11), implies that $z^{**} = g(z^*; \Gamma)$ or, equivalently, $z(\Delta t) = z^{**}$ if $z(0) = z^*$ and

$$\{(r, \tau_{i1}(r), \cdots, \tau_{in}(r)) \mid r \in \mathcal{T}_i \cap \{0, \cdots, \Delta t\}\} = \Gamma_i, \quad \forall i.$$

Since $d(z^*) = \beta > 0$, this, together with Lemma 2.3(c), implies that $d(z^{**}) < d(z^*)$, contradicting the hypothesis $d(z^{**}) = \beta$. □

The convexity of X^* is sometimes hard to verify. For this reason we will consider another assumption that is stronger than Assumption B but is easier to verify.

Assumption B'.

(a) f is continuous.

(b) The set of fixed points X^* is nonempty.

(c) $\|f(x)-x^*\| \leqq \|x-x^*\|$, for all $x \in \Re^n$, for all $x^* \in X^*$.

(d) For every $x \notin X^*$ and $x^* \in X^*$, there exists some $i \in I(x; x^*)$ such that $f_i(y) \neq y_i$ for all $y \in U(x; x^*)$ such that $y \notin X^*$.

Compared to Assumption B, part (d) of the new assumption is stronger but part (b) is weaker because convexity is not assumed. We have the following result.

LEMMA 2.4. *Assumption* B' *implies Assumption* B.

Proof. It can be seen that Assumption B'(d) implies Assumption B(d), so we only need to show that X^* is convex. Suppose the contrary. Since X^* is closed, then there exist $x^* \in X^*$ and $y^* \in X^*$ such that $(x^*+y^*)/2 \notin X^*$. Let $x=(x^*+y^*)/2$. It can be seen that $\|x-x^*\| = \|x-y^*\| > 0$, $x \notin X^*$, and $I(x; x^*) = I(x; y^*)$ (see Fig. 2.3). By Assumption B'(d), there exists $i \in I(x; x^*)$ such that $f_i(x) \neq x_i$. Suppose that $x_i > y_i^*$. Then if $f_i(x) > x_i$, we obtain $\|f(x)-y^*\| \geqq f_i(x)-y_i^* > x_i - y_i^* = \|x-y^*\|$ and if $f_i(x) < x_i$, we similarly obtain $\|f(x)-x^*\| > \|x-x^*\|$. In either case Assumption B'(c) is contradicted. The case where $x_i < y_i^*$ is treated analogously. □

Assumption B will be used in § 4, while Assumption B' will be used in §§ 3, 6, and 7.

3. Nonexpansive mappings on a box. Let $g: \Re^n \to \Re^n$ be a continuously differentiable function satisfying the following assumption:

Assumption D.

(a) For each i, $\sum_{j=1}^n |\partial g_i(x)/\partial x_j| \leqq 1$, for all $x \in \Re^n$.

(b) For each i and j, either $\partial g_i(x)/\partial x_j = 0$, for all $x \in \Re^n$, or $\partial g_i(x)/\partial x_j \neq 0$, for all $x \in \Re^n$.

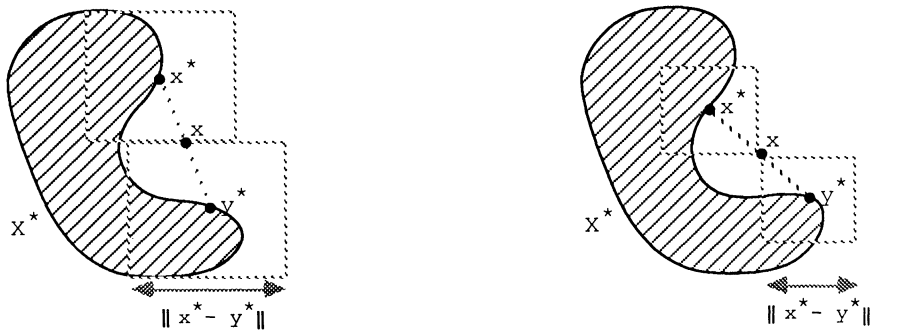


FIG. 2.3. *Two configurations of* x^* *and* y^*.

(c) The graph with node set $\{1, \cdots, n\}$ and arc set $\{(i, j) \mid \partial g_i(x)/\partial x_j \neq 0\}$ is strongly connected.

Let C be a box (possibly unbounded) in $\Re^n$, i.e.,

$$C = \{x \in \Re^n \mid l_i \leqq x_i \leqq c_i, \ \forall i\},$$

for some scalars l_i and c_i satisfying $l_i \leqq c_i$ (we allow $l_i = -\infty$ or $c_i = +\infty$). Let also $[x]^+$

denote the orthogonal projection of x onto C, i.e.,

$$[x]^+ = (\max\{l_1, \min\{c_1, x_1\}\}, \cdots, \max\{l_n, \min\{c_n, x_n\}\}).$$

We use the notation x^T to denote the transpose of a column vector x. The following is the main result of this section.

PROPOSITION 3.1. *Let $g: \Re^n \to \Re^n$ satisfy Assumption* D. *If either g has a fixed point or if C is bounded, then the function $h: \Re^n \to \Re^n$ defined by*

$$h(x) = [g(x)]^+ \tag{3.1}$$

satisfies Assumption B′.

Proof. Since both g and $[\cdot]^+$ are continuous functions, so is their composition, and part (a) of Assumption B′ holds.

By the Mean Value Theorem, for any $x \in \Re^n$, $y \in \Re^n$, and index i, there exists $\xi \in \Re^n$ such that

$$g_i(y) - g_i(x) = (\nabla g_i(\xi))^T (y - x). \tag{3.2}$$

This implies that

$$\begin{aligned} |g_i(y) - g_i(x)| &\leqq \sum_j |\partial g_i(\xi)/\partial x_j| |y_i - x_i| \\ &\leqq \left(\sum_j |\partial g_i(\xi)/\partial x_j| \right) \|x - y\| \\ &\leqq \|x - y\|, \end{aligned}$$

where the last inequality follows from Assumption D(a). Since the choice of i was arbitrary, g is nonexpansive with respect to the maximum norm. Since projection onto a box can be easily seen as nonexpansive with respect to the maximum norm, it follows from (3.1) that $\|h(x) - h(y)\| \leqq \|g(x) - g(y)\|$. Thus, h is nonexpansive with respect to the maximum norm, and part (c) of Assumption B′ is satisfied.

We now show that h has a fixed point. Suppose first that g has a fixed point y^*. Choose β sufficiently large so that the set $Y = \{x \in \Re^n \mid \|x - y^*\| \leqq \beta\} \cap C$ is nonempty. Then for every $x \in Y$ we have, for all i,

$$y_i^* - \beta \leqq g_i(x) \leqq y_i^* + \beta,$$

and

$$\text{either} \quad l_i \leqq g_i(x) \leqq c_i \quad \text{or} \quad g_i(x) < l_i \leqq y_i^* + \beta \quad \text{or} \quad y_i^* - \beta \leqq c_i < g_i(x).$$

Since $h_i(x) = \max\{l_i, \min\{c_i, g_i(x)\}\}$, this implies that $h(x) \in Y$ (see Fig. 3.1 below).

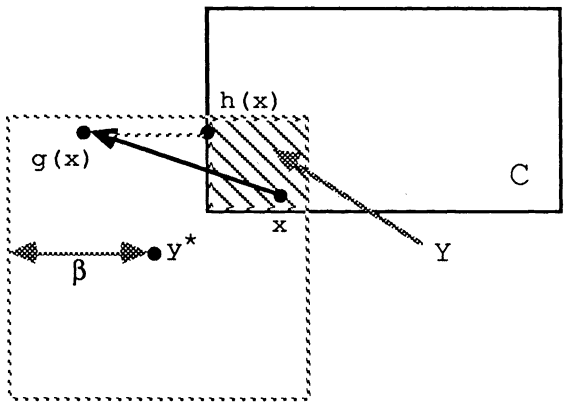


FIG. 3.1

Since h is also continuous and Y is convex and compact, a theorem of Brouwer ([11], p. 17) shows that h has a fixed point. Now suppose C is bounded. Since $h(x) \in C$ for all $x \in C$ and C is convex and compact, the same theorem of Brouwer shows that h has a fixed point. Thus, part (b) of Assumption B′ is satisfied.

We finally show that Assumption B′(d) holds. Suppose the contrary. Then there exists some $x \notin X^*$ and some $x^* \in X^*$, such that for every $i \in I(x; x^*)$ there is an $x^i \in U(x; x^*)$ with $x^i \notin X^*$ and $h_i(x^i) = x_i^i$. Let $J = I(x; x^*)$, $\beta = \|x - x^*\|$ and fix some $i \in J$. By the Mean Value Theorem, there exists some $\xi \in \Re^n$ such that $g_i(x^i) - g_i(x^*) = (\nabla g_i(\xi))^T (x^i - x^*)$. Let $a_j = \partial g_i(\xi)/\partial x_j$. Then

$$\begin{aligned}
\beta &= |x_i^i - x_i^*| = |h_i(x^i) - h_i(x^*)| \\
&\leqq |g_i(x^i) - g_i(x^*)| \\
&= \left| \sum_j a_j (x_j^i - x_j^*) \right| \\
&\leqq \left(\sum_{j \in J} |a_j| \right) \beta + \left(\sum_{j \notin J} |a_j| |x_j^i - x_j^*| \right) \\
&\leqq \beta + \sum_{j \notin J} |a_j| (|x_j^i - x_j^*| - \beta),
\end{aligned}$$

where the first inequality follows from the fact that the projection onto $[l_i, c_i]$ is nonexpansive and the last inequality follows from the fact (cf. Assumption D(a)) that $\sum_j |a_j| \leqq 1$. Since $|x_j^i - x_j^*| < \beta$ for all $j \notin J$, the above inequality implies that $a_j = 0$ for all $j \notin J$. Since the choice of $i \in J$ was arbitrary, we obtain from Assumption D(b) that $\partial g_i(\xi)/\partial x_j = 0$ for all $\xi \in \Re^n$, $i \in J$, $j \notin J$. By Assumption D(c), we must have that $J = \{1, \cdots, n\}$. In that case, $U(x; x^*)$ is a singleton and all the vectors x^i are equal. It then follows from the equalities $h_i(x^i) = x_i^i$, for all i, that each x^i is a fixed point of h, a contradiction of the hypothesis $x^i \notin X^*$. □

Since Assumption B′ is satisfied, the partially asynchronous iteration

$$x := (1 - \gamma)x + \gamma [g(x)]^+$$

(with $0 < \gamma < 1$) converges (cf. Lemmas 2.1, 2.4, and Proposition 2.1).

An important special case is obtained if $C = \Re^n$, $g(x) = Ax + b$, where A is an $n \times n$ matrix and b is a given vector in $\Re^n$. Thus, the problem is to solve the linear system

$$x = Ax + b,$$

and Assumption D amounts to the requirement that $A = [a_{ij}]$ is irreducible (see [22] for a definition of irreducibility) and $\sum_j |a_{ij}| \leqq 1$, for all i. Then, provided that the system $x = Ax + b$ has a solution (not necessarily unique), the partially asynchronous iteration

$$x := (1 - \gamma)x + \gamma(Ax + b)$$

(with $0 < \gamma < 1$) will converge to such a solution.

As a special case of our results, we obtain convergence of the synchronous iteration

$$x(t+1) = (1 - \gamma)x(t) + \gamma(Ax(t) + b).$$

This seems to be a new result under our assumptions. Previous convergence results [17], [22] have made the stronger assumption that either: (a) A is irreducible and $\sum_j |a_{ij}| \leqq 1$, for all i, with strict inequality for at least one i, or (b) $\sum_j |a_{ij}| < 1$, for all i. Two other important special cases are studied below.

3.1. Quadratic costs subject to box constraints. Consider the following problem.

$$\begin{aligned} &\text{Minimize} && x^TQx/2+p^Tx \\ &\text{Subject to} && x\in C, \end{aligned} \tag{3.3}$$

where $Q=[q_{ij}]$ is a symmetric, irreducible, nonnegative definite matrix of dimension $n\times n$ satisfying the weak diagonal dominance condition

$$\sum_{j\neq i}|q_{ij}|\leqq q_{ii},\quad q_{ii}>0,\quad \forall i, \tag{3.4}$$

p is an element of $\Re^n$, and C is, as before, a box in $\Re^n$.

Let D denote the diagonal matrix whose ith diagonal entry is q_{ii}. Let $A=I-D^{-1}Q$ and $b=-D^{-1}p$. We have the following result.

PROPOSITION 3.2. *The function $g:\Re^n\to\Re^n$ defined by $g(x)=Ax+b$ satisfies Assumption* D.

Proof. g is clearly continuously differentiable and (cf. (3.4)) $\sum_j|a_{ij}|=\sum_{j\neq i}|q_{ij}|/q_{ii}\leqq 1$ for all i. Since $\partial g_i(x)/\partial x_j=a_{ij}$ for all $x\in\Re^n$ and A is irreducible, g satisfies Assumption D. □

It can be seen (by using the Kuhn-Tucker optimality conditions [23]) that each optimal solution of (3.3) is a fixed point of $[Ax+b]^+$ and vice versa, where $[\cdot]^+$ denotes the orthogonal projection onto C. Hence, if (3.3) has an optimal solution, then (cf. Lemma 2.1, 2.4, and Propositions 2.1, 3.1, 3.2) the partially asynchronous iteration

$$x:=(1-\gamma)x+\gamma[Ax+b]^+ \tag{3.5}$$

(with $0<\gamma<1$) converges to such a solution. Note that for $\gamma=1$, the iteration (3.5) takes the form $x:=[x-D^{-1}(Qx+p)]^+$ which is a diagonally scaled gradient projection iteration. However, this iteration need not be convergent in the absence of additional assumptions.

3.2. Separable quadratic costs with sparse 0, +1, −1 matrix. Consider the following problem.

$$\begin{aligned} &\text{Minimize} && w^TDw/2+\beta^Tw \\ &\text{Subject to} && Ew\geqq d, \end{aligned} \tag{3.6}$$

where D is an $m\times m$ positive definite diagonal matrix, β is an element of $\Re^m$, d is an element of $\Re^n$, and $E=[e_{ik}]$ is an $n\times m$ matrix having at most two nonzero entries per column, and each nonzero entry is either -1 or 1. Furthermore, we assume that the undirected graph $\mathcal{G}$ with node set $\{1,\cdots,n\}$ and arc set $\{(i,j)\mid e_{ik}\neq 0$ and $e_{jk}\neq 0$ for some $k\}$ is connected.

Consider the following Lagrangian dual [23] of (3.6).

$$\begin{aligned} &\text{Minimize} && x^TQx/2+p^Tx \\ &\text{Subject to} && x\geqq 0, \end{aligned}$$

where $Q=ED^{-1}E^T$, $p=-d-ED^{-1}\beta$. We show below that this is a special case of the problem considered in the previous subsection.

PROPOSITION 3.3. *Q is symmetric, irreducible, nonnegative definite and weakly diagonally dominant* (cf. (3.4)).

Proof. Since D is symmetric and positive definite, Q is symmetric and nonnegative definite. To see that Q satisfies (3.4), let α_k denote the kth diagonal entry of D ($\alpha_k>0$), let $O(i)$ denote the set of indices k such that $e_{ik}\neq 0$, and let q_{ij} denote the (i,j)th entry

of Q. Then

$$|q_{ij}| = \left|\sum_k e_{ik}(\alpha_k)^{-1}e_{jk}\right| \leqq \sum_{k\in O(i)\cap O(j)} (\alpha_k)^{-1},$$

with equality holding if $i=j$. Hence, for each i,

$$\sum_{j\neq i}|q_{ij}| \leqq \sum_{j\neq i}\sum_{k\in O(i)\cap O(j)}(\alpha_k)^{-1} \leqq \sum_{k\in O(i)}(\alpha_k)^{-1} = q_{ii},$$

where the second inequality follows from the fact that if $k\in O(i)\cap O(j)$ for some j, then $k\notin O(i)\cap O(j')$ for all j' not equal to i or j. Finally, Q is irreducible because $\mathcal{G}$ is connected and $q_{ij}\neq 0$ for $i\neq j$ if and only if there exists some k such that $e_{ik}\neq 0$ and $e_{jk}\neq 0$. □

An example of constraints $Ew\geqq d$ satisfying our conditions on E is

$$\sum_k w_k \leqq 1 \quad \text{and} \quad \sum_{k\in K_r} w_k \geqq 0 \quad \text{for } r=1,2,\cdots,R,$$

where $K_1, K_2, \cdots, K_R$ are some mutually disjoint subsets of $\{1,2,\cdots,m\}$. Such constraints often arise in resource allocation problems.

4. Strictly convex cost network flow problems. Consider a connected, directed graph (network) with the set of nodes $\mathcal{N}=\{1,\cdots,n\}$ and the set of arcs $\mathcal{A}\subseteq\mathcal{N}\times\mathcal{N}$. We assume that $i\neq j$ for every arc (i,j) and that at most one arc connects any ordered pair of nodes, so that the arc (i,j) has unambiguous meaning. (These restrictions can be easily removed.) For each node $i\in\mathcal{N}$, denote by $\mathcal{D}(i)$ the set of downstream neighbors of i (that is, $\mathcal{D}(i)=\{j\,|\,(i,j)\in\mathcal{A}\}$) and by $\mathcal{U}(i)$ the set of upstream neighbors of i (that is, $U(i)=\{j\,|\,(j,i)\in\mathcal{A}\}$). Consider the following problem:

$$\text{(4.1)} \qquad \text{Minimize} \quad \sum_{(i,j)\in\mathcal{A}} a_{ij}(f_{ij})$$

$$\text{(4.2)} \qquad \text{Subject to} \quad \sum_{j\in\mathcal{D}(i)} f_{ij} - \sum_{j\in\mathcal{U}(i)} f_{ji} = s_i, \quad \forall i\in\mathcal{N},$$

where each $a_{ij}:\Re\to(-\infty,+\infty]$ is a strictly convex, lower semicontinuous function and each s_i is a real number. We interpret f_{ij} as the *flow* on the arc (i,j), s_i as the *supply* (or *demand* if $s_i<0$) at node i, and $a_{ij}(f_{ij})$ as the *cost* of sending a flow of f_{ij} on arc (i,j). The goal is then to find a set of arc flows that minimizes the total cost while satisfying the flow conservation constraints (4.2) (see Fig. 4.1). Note that capacity constraints of the form

$$b_{ij}\leqq f_{ij}\leqq c_{ij},$$

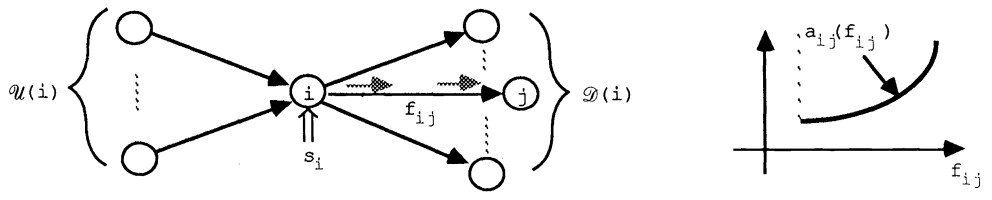


FIG. 4.1

where b_{ij}, c_{ij} are given scalars, can be incorporated into the cost function a_{ij} by letting $a_{ij}(f_{ij}) = +\infty$ for $f_{ij} \notin [b_{ij}, c_{ij}]$.

The above network flow problem is an important optimization problem, with applications to data networks, traffic assignment, matrix balancing, etc. The interested reader is referred to [7, Chap. 5] for a detailed discussion of this problem. (Also see [5], [6], [9], [12], [21], [24], [31]-[33].)

Denote by $g_{ij}: \Re \to (-\infty, +\infty]$ the *conjugate function* ([23, § 12]; [24, p. 330]) of a_{ij}, i.e.,

$$g_{ij}(\eta) = \sup_{\zeta \in \Re} \{\zeta\eta - a_{ij}(\zeta)\}. \tag{4.3}$$

Each g_{ij} is convex and, by assigning a Lagrange multiplier p_i (also called a *price*) to the ith constraint of (4.2), we can formulate the dual problem ([24, § 8G]) of (4.1) as the following convex minimization problem.

$$\begin{aligned} &\text{Minimize} \quad && q(p) = \sum_{(i,j) \in \mathcal{A}} g_{ij}(p_i - p_j) - \sum_{i \in \mathcal{N}} p_i s_i \\ &\text{Subject to} \quad && p \in \Re^n. \end{aligned} \tag{4.4}$$

We make the following assumption.

Assumption E.

(a) Each conjugate function g_{ij} is real valued.

(b) The set P^* of optimal solutions of the dual problem (4.4) is nonempty.

Assumption E implies (cf. [24, § 11D]) that the original problem (4.1) has an optimal solution, and the optimal objective value for (4.1) and (4.4) sum to zero. Furthermore, the strict convexity of the a_{ij}'s implies that (4.1) has a *unique* optimal solution, which we denote by $f^* = (\cdots, f_{ij}^*, \cdots)_{(i,j) \in \mathcal{A}}$, and that every g_{ij} is continuously differentiable ([23, pp. 218, 253]). Hence q given by (4.4) is also continuously differentiable. Its partial derivative $\partial q(p)/\partial p_i$, to be denoted by $d_i(p)$, is given by

$$d_i(p) = \frac{\partial q(p)}{\partial p_i} = \sum_{j \in \mathcal{D}(i)} \nabla g_{ij}(p_i - p_j) - \sum_{j \in \mathcal{U}(i)} \nabla g_{ji}(p_j - p_i) - s_i. \tag{4.5}$$

Given a price vector $p \in \Re^n$, we consider an iteration whereby the dual objective function q is minimized with respect to the ith coordinate p_i, while the remaining coordinates are held fixed. In view of the convexity and the differentiability of q, this is equivalent to solving the equation $d_i(p_1, \cdots, p_{i-1}, \theta, p_{i+1}, \cdots, p_n) = 0$ with respect to the scalar θ. This equation can have several solutions and we will consider a mapping which chooses the solution that is nearest to the original price p_i. Accordingly, we define a function $h: \Re^n \to \Re^n$ whose ith coordinate is given by

$$h_i(p) = \operatorname{argmin} \{|\theta - p_i| \mid d_i(p_1, \cdots, p_{i-1}, \theta, p_{i+1}, \cdots, p_n) = 0\}. \tag{4.6}$$

We will show later in Lemma 4.1 that the set in (4.6) is nonempty and the minimum in (4.6) is attained, so that h is well defined. Notice that $h(p) = p$ if and only if $\partial q(p)/\partial p_i = d_i(p) = 0$ for every i. It follows that P^* is the set of fixed points of h.

Since q is convex, the set P^* is convex (P^* is also nonempty by assumption). Also from Proposition 2.3 in [6] we have that, for any $p \in \Re^n$ and any $p^* \in P^*$,

$$\min_{j \in \mathcal{N}} \{p_j - p_j^*\} \leqq h_i(p) - p_i^* \leqq \max_{j \in \mathcal{N}} \{p_j - p_j^*\}, \quad \forall i \in \mathcal{N},$$

and hence h has the pseudo-nonexpansive property

$$\|h(p) - p^*\| \leqq \|p - p^*\|.$$

Furthermore, by using Proposition 1 in [5] and an argument analogous to the proof of Proposition 2.5 in Chapter 7.2 of [7], we can show that the mapping h is continuous. Therefore, h satisfies parts (a)-(c) of Assumption B. We show below that h is well defined and also satisfies part (d) of Assumption B.

LEMMA 4.1. *The mapping h is well defined and satisfies Assumption* B(d).

Proof. We start by mentioning certain facts that will be freely used in the course of the proof.

(a) For any $(i, j) \in \mathcal{A}$, the function ∇g_{ij} is nondecreasing. (This is because g_{ij} is convex.)

(b) $d_i : \Re^n \to \Re$ is a nondecreasing function of the ith coordinate of its argument when the other coordinates are held fixed. (This is because the dual functional q is convex and $d_i = \partial q / \partial p_i$.)

(c) A vector $p^* \in \Re^n$ belongs to P^* if and only if, for every arc (i, j), we have $\nabla g_{ij}(p_i^* - p_j^*) = f_{ij}^*$. (This is a direct consequence of the Network Equilibrium Theorem in [24, p. 349].)

We first show that h is well defined. Fix any $p \in \Re^n$ and any i. We claim that there exists θ_1 such that $d_i(p + \theta_1 e^i) \leqq 0$, where e^i denotes the ith coordinate vector in $\Re^n$. To see this, let p^* be any element of P^* and let θ_1 be any scalar sufficiently large so that

$$p_i - p_j + \theta_1 \geqq p_i^* - p_j^*, \quad \forall j \in \mathcal{D}(i),$$
$$p_j - p_i - \theta_1 \leqq p_j^* - p_i^*, \quad \forall j \in \mathcal{U}(i).$$

Since ∇g_{kl} is nondecreasing for all $(k, l) \in \mathcal{A}$, this implies that

$$\nabla g_{ij}(p_i - p_j + \theta_1) \geqq \nabla g_{ij}(p_i^* - p_j^*) = f_{ij}^*, \quad \forall j \in \mathcal{D}(i),$$
$$\nabla g_{ji}(p_j - p_i - \theta_1) \leqq \nabla g_{ji}(p_j^* - p_i^*) = f_{ji}^*, \quad \forall j \in \mathcal{U}(i).$$

Upon summing the above inequalities, we obtain that

$$\begin{aligned} d_i(p + \theta_1 e^i) &= \sum_{j \in \mathcal{D}(i)} \nabla g_{ij}(p_i - p_j + \theta_1) - \sum_{j \in \mathcal{U}(i)} \nabla g_{ji}(p_j - p_i - \theta_1) - s_i \\ &\geqq \sum_{j \in \mathcal{D}(i)} f_{ij}^* - \sum_{j \in \mathcal{U}(i)} f_{ji}^* - s_i \\ &= 0, \end{aligned}$$

where the last equality follows because the flows f_{ij}^* and f_{ji}^* must satisfy the flow conservation equation (4.2). Similarly, we can show that there exists θ_2 such that $d_i(p + \theta_2 e^i) \leqq 0$. Since $d_i(p + \theta e^i)$ is a continuous function of θ, this implies that there exists some θ between θ_1 and θ_2 such that $d_i(p + \theta e^i) = 0$. Therefore the set in (4.6) is nonempty. Since this set is also convex (due to the convexity of q) and closed (due to the continuity of d_i), the minimum in (4.6) is attained. Hence h is well defined.

Now we show that h satisfies Assumption B(d). We will argue by contradiction. Suppose that h does not satisfy Assumption B(d). Then for some $p \notin P^*$ and $p^* \notin P^*$ such that $\|p - p^*\| = \rho(p) > 0$ there exists, for every $i \in I(p; p^*)$, a vector $p^i \in U(p; p^*)$ such that $h_i(p^i) = p_i^i$. ($\rho(p)$ denotes the maximum norm distance of p from P^*.) Let $\beta = \rho(p)$, $J = I(p; p^*)$, $\varepsilon = \beta - \max\{|p_i^k - p_i^*| \mid i \notin J, k \in J\}$, and

$$J^- = \{i \mid p_i - p_i^* = -\beta\},$$
$$J^+ = \{i \mid p_i - p_i^* = \beta\}.$$

Then $\varepsilon > 0$, $J = J^- \cup J^+$ and, for all $i \in J$,

$$p_j^* - \beta + \varepsilon \leqq p_j^i \leqq p_j^* + \beta - \varepsilon, \quad \forall j \notin J, \tag{4.7}$$
$$p_j^i = p_j^* - \beta, \quad \forall j \in J^-, \tag{4.8a}$$
$$p_j^i = p_j^* + \beta, \quad \forall j \in J^+. \tag{4.8b}$$

Fix any $i \in J^-$. The relations (4.7), (4.8a) imply that

$$p_i^i - p_j^i \leqq (p_i^* - \beta) - (p_j^* - \beta) = p_i^* - p_j^*, \quad \forall j \in \mathcal{D}(i),$$
$$p_j^i - p_i^i \geqq (p_j^* - \beta) - (p_i^* - \beta) = p_j^* - p_i^*, \quad \forall j \in \mathcal{U}(i),$$

and, since ∇g_{kl} is nondecreasing for all $(k, l) \in \mathcal{A}$,

$$\nabla g_{ij}(p_i^i - p_j^i) \leqq \nabla g_{ij}(p_i^* - p_j^*) = f_{ij}^*, \quad \forall j \in \mathcal{D}(i), \tag{4.9a}$$

$$\nabla g_{ji}(p_j^i - p_i^i) \geqq \nabla g_{ji}(p_j^* - p_i^*) = f_{ji}^*, \quad \forall j \in \mathcal{U}(i). \tag{4.9b}$$

Since $i \in J^-$, we have $h_i(p^i) = p_i^i$ or, equivalently, $d_i(p^i) = 0$. Then (4.5) and (4.9a)-(4.9b) imply that

$$\begin{aligned} 0 &= d_i(p^i) \\ &= \sum_{j \in \mathcal{D}(i)} \nabla g_{ij}(p_i^i - p_j^i) - \sum_{j \in \mathcal{U}(i)} \nabla g_{ji}(p_j^i - p_i^i) - s_i \\ &\leqq \sum_{j \in \mathcal{D}(i)} f_{ij}^* - \sum_{j \in \mathcal{U}(i)} f_{ji}^* - s_i \\ &= 0, \end{aligned}$$

where the last equality follows because the flows f_{ij}^* and f_{ji}^* must satisfy the flow conservation equation (4.2). It follows that the inequalities in (4.9a)-(4.9b) are actually equalities and

$$\nabla g_{ij}(p_i^i - p_j^i) = f_{ij}^*, \quad \forall j \in \mathcal{D}(i), \tag{4.10a}$$

$$\nabla g_{ji}(p_j^i - p_i^i) = f_{ji}^*, \quad \forall j \in \mathcal{U}(i). \tag{4.10b}$$

Since the choice of $i \in J^-$ was arbitrary, (4.10a)-(4.10b) hold for all $i \in J^-$. By an analogous argument (using (4.8b) in place of (4.8a)) we can show that (4.10a)-(4.10b) hold for all $i \in J^+$ as well.

Let $\pi \in \Re^n$ be the vector whose ith component is

$$\pi_i = \begin{cases} p_i^* + \varepsilon & \text{if } i \in J^+, \\ p_i^* - \varepsilon & \text{if } i \in J^-, \\ p_i^* & \text{if } i \notin J. \end{cases} \tag{4.11}$$

We claim that

$$\nabla g_{ij}(\pi_i - \pi_j) = f_{ij}^*, \quad \forall (i, j) \in \mathcal{A}. \tag{4.12}$$

To see this, we first note from the definition of π (cf. (4.11)) that

$$\pi_i - \pi_j = p_i^* - p_j^*, \quad \text{if } i \notin J, j \notin J \quad \text{or if } i \in J^+, j \in J^+ \quad \text{or if } i \in J^-, j \in J^-.$$

Also, from (4.7), (4.8a)-(4.8b), (4.11) and the fact $\varepsilon \leqq \beta$ we have that

$$\begin{aligned}
&p_i^i - p_j^i = (p_i^* + \beta) - (p_j^* - \beta) \geqq \pi_i - \pi_j \geqq p_i^* - p_j^*, && \text{if } i \in J^+, j \in J^-, \\
&p_i^i - p_j^i = (p_i^* - \beta) - (p_j^* + \beta) \leqq \pi_i - \pi_j \leqq p_i^* - p_j^*, && \text{if } i \in J^-, j \in J^+, \\
&p_i^i - p_j^i \geqq (p_i^* + \beta) - (p_j^* + \beta - \varepsilon) = \pi_i - \pi_j \geqq p_i^* - p_j^*, && \text{if } i \in J^+, j \notin J, \\
&p_i^i - p_j^i \leqq (p_i^* - \beta) - (p_j^* - \beta + \varepsilon) = \pi_i - \pi_j \leqq p_i^* - p_j^*, && \text{if } i \in J^-, j \notin J, \\
&p_i^i - p_j^i \leqq (p_i^* + \beta - \varepsilon) - (p_j^* + \beta) = \pi_i - \pi_j \leqq p_i^* - p_j^*, && \text{if } i \notin J, j \in J^+, \\
&p_i^i - p_j^i \geqq (p_i^* - \beta + \varepsilon) - (p_j^* - \beta) = \pi_i - \pi_j \geqq p_i^* - p_j^*, && \text{if } i \notin J, j \in J^-.
\end{aligned}$$

Consider any $(i, j) \in \mathcal{A}$. The preceding inequalities show that $\pi_i - \pi_j$ is always between $p_i^i - p_j^i$ and $p_i^* - p_j^*$. The monotonicity of ∇g_{ij} and the equalities $\nabla g_{ij}(p_i^* - p_j^*) = f_{ij}^* = \nabla g_{ij}(p_i^i - p_j^i)$ (cf. (4.10a)-(4.10b)) imply that $\nabla g_{ij}(\pi_i - \pi_j) = f_{ij}^*$. This completes the proof of (4.12).

Equation (4.12) implies that $\pi \in P^*$. Since (cf. (4.11) and the definitions of J^- and J^+) $\|p - \pi\| < \|p - p^*\|$, this contradicts the hypothesis that $\rho(p) = \|p - p^*\|$. □

Since h has been shown to satisfy Assumption B, we conclude from Lemma 2.1 and Proposition 2.1 that the partially asynchronous iteration

$$p := (1-\gamma)p + \gamma h(p)$$

(with $0 < \gamma < 1$) converges to an optimal price vector p^*. The optimal flows are obtained as a byproduct, using the relation $\nabla g_{ij}(p_i^* - p_j^*) = f_{ij}^*$. Notice that the iteration for each coordinate p_i consists of minimization along the ith coordinate direction (to obtain $h_i(p)$) followed by the use of the relaxation parameter γ to obtain the new value $(1-\gamma)p_i + \gamma h_i(p)$. As a special case, we have that the synchronous Jacobi algorithm

$$p(t+1) = (1-\gamma)p(t) + \gamma h(p(t))$$

is also convergent, which is a new result.

A related result can be found in [5] where totally asynchronous convergence is established even if $\gamma = 1$, provided that a particular coordinate of p is never iterated upon and that when this coordinate is fixed, the optimal price vector is unique. An experimental comparison of the two methods will be presented in § 8. We remark that the results in this section also extend to the case where each arc has a gain of either $+1$ or -1 (i.e., each f_{ji} term in (4.2) is multiplied by either $+1$ or -1).

5. Agreement and Markov chain algorithms. In this section we consider two problems: a problem of agreement and the computation of the invariant distribution of a Markov chain. These problems are the only ones for which partially asynchronous algorithms that converge for every value of the asynchronism bound B of Assumption A are available [20], [27], [29] (in fact, these algorithms have been shown to converge at a geometric rate). We show that these results can also be obtained by applying our general convergence theorem (Proposition 2.1).

5.1. The agreement algorithm. We consider here a set of n processors, numbered from 1 to n, that try to reach agreement on a common value by exchanging tentative values and forming convex combinations of their own values with the values received from other processors. This algorithm has been used in [28]-[29] in the context of asynchronous stochastic gradient methods with the purpose of averaging noisy measurements of the same variable by different processors.

We now formally describe the agreement algorithm. Each processor i has a set of nonnegative coefficients $\{a_{i1}, \cdots, a_{in}\}$ satisfying $a_{ii} > 0$, $\sum_j a_{ij} = 1$, and at time t it possesses an estimate $x_i(t)$ which is updated according to (cf. (1.1))

$$x_i(t+1) = \begin{cases} \sum_{j=1}^{n} a_{ij} x_j(\tau_{ij}(t)) & \text{if } t \in \mathcal{T}_i, \\ x_i(t) & \text{otherwise.} \end{cases} \tag{5.1a}$$

$$x_i(1-B) = \cdots = x_i(0) = \bar{x}_i, \tag{5.1b}$$

where $\mathcal{T}_i$ and $\tau_{ij}(t)$ are as in § 1 and $\bar{x}_i$ is the initial value of processor i. Let A be the $n \times n$ matrix whose (i, j)th entry is a_{ij} and let $\gamma \in (0, 1)$ be such that $0 < \gamma \leqq \min\{a_{11}, \cdots, a_{nn}\}$. By using the results from §§ 1 to 3 we obtain the following.

PROPOSITION 5.1. *If A is irreducible and Assumption* A *holds, then $\{x_i(t)\} \to y$ for all i, where y is some scalar between* $\min_i \{\bar{x}_i\}$ *and* $\max_i \{\bar{x}_i\}$.

Proof. It can be seen that (5.1a) is a special case of (1.1) with $f(x) = Ax$. Let

$$D = (A - \gamma I)/(1 - \gamma).$$

Then

$$A = \gamma I + (1 - \gamma) D, \tag{5.2}$$

and $D = [d_{ij}]$ can be seen to satisfy $\sum_j |d_{ij}| \leqq 1$. Moreover, since A is irreducible, so is D. Hence the function $h : \Re^n \to \Re^n$ defined by $h(x) = Dx$ satisfies Assumption D in § 3. Since h has a fixed point (the zero vector), this, together with Proposition 3.1 and Lemma 2.4, implies that h satisfies Assumption B. Since (cf. (5.2)) $f(x) = \gamma x + (1 - \gamma) h(x)$, this, together with Lemma 2.1, shows that f satisfies Assumption C. Then by Proposition 2.1, the sequence $\{x(t)\}$ generated by (5.1a)-(5.1b) converges to some point x^∞ satisfying $Ax^\infty = x^\infty$. Since A is irreducible and stochastic, x^∞ must be of the form $(y, \cdots, y)$ for some $y \in \Re$. It can be seen from (5.1b) that, for $r \in \{1 - B, \cdots, 0\}$,

$$x_i(r) \leqq \max_j \{\bar{x}_j\}, \quad \forall i. \tag{5.3}$$

Suppose that (5.3) holds for all $r \in \{1 - B, \cdots, t\}$, for some $t \geqq 0$. Then by (5.1a) and the property of the a_{ij}'s,

$$\begin{aligned} x_i(t+1) &= \sum_j a_{ij} x_j(\tau_{ij}(t)) \\ &\leqq \sum_j a_{ij} \max_j \{\bar{x}_j\} \\ &= \max_j \{\bar{x}_j\}, \end{aligned}$$

for all i such that $t \in \mathcal{T}_i$, and $x_i(t+1) = x_i(t) \leqq \max_j \{\bar{x}_j\}$ for all other i. Hence, by induction, (5.3) holds for all $r \in \{1 - B, 2 - B, \cdots\}$. Since $x_i(r) \to y$ for each i, this implies that $y \leqq \max_j \{\bar{x}_j\}$. A symmetrical argument shows $y \geqq \min_j \{\bar{x}_j\}$. □

It can be shown [7], [29] that Proposition 5.1 remains valid if a_{ii} is positive for at least one (but not all) i and, furthermore, convergence takes place at the rate of a geometric progression. The proof, however, is more complex. Similar results can be found in [29] for more general versions of the agreement algorithm.

5.2. Invariant distribution of Markov chains. Let P be an irreducible stochastic matrix of dimension $n \times n$. We denote by p_{ij} the (i, j)th entry of P and we assume that $p_{ii} > 0$ for all i. We wish to compute a row vector $\pi^* = (\pi_1^*, \cdots, \pi_n^*)$ of invariant probabilities for the corresponding Markov chain, i.e., $\pi_i^* \geqq 0$, $\sum_i \pi_i^* = 1$, $\pi^* = \pi^* P$. (We actually have $\pi_i^* > 0$, for all i, due to the irreducibility of P [14].) As in § 5.1, suppose that we have a network of n processors and that the ith processor generates a sequence of estimates $\{\pi_i(t)\}$ using the following partially asynchronous version of the classical serial algorithm $\pi := \pi P$ (cf. (5.1a)-(5.1b)):

$$\pi_i(t+1) = \begin{cases} \sum_{j=1}^{n} p_{ji} \pi_j(\tau_{ij}(t)) & \text{if } t \in \mathcal{T}_i, \\ \pi_i(t) & \text{otherwise.} \end{cases} \tag{5.4}$$

$$\pi_i(1 - B) = \cdots = \pi_i(0),$$

where $\mathcal{T}_i$ and $\tau_{ij}(t)$ are as in § 1 and $\pi_i(0)$ is any positive scalar. This asynchronous algorithm was introduced in [20], where geometric convergence was established. We show below that convergence also follows from our general results.

PROPOSITION 5.2. *If Assumption* A *holds, then there exists a positive number c such that* $\pi(t) \to c\pi^*$.

Proof. We will show that (5.4) is a special case of (5.1a). Let

$$x_i(t) = \pi_i(t)/\pi_i^*, \qquad a_{ij} = \pi_j^* p_{ji}/\pi_i^*. \tag{5.5}$$

Then the matrix $A = [a_{ij}]$ is nonnegative and irreducible, has positive diagonal entries, and

$$\begin{aligned} \sum_j a_{ij} &= \sum_j \pi_j^* p_{ji}/\pi_i^* \\ &= \pi_i^*/\pi_i^* \\ &= 1, \end{aligned}$$

where the second equality follows from $\pi^* = \pi^* P$. Furthermore, it can be seen from (5.4) and (5.5) that $x_i(t)$ evolves according to the iteration (5.1a). Therefore, by Proposition 5.1 and the initial positivity conditions, $\{x_i(t)\} \to c$ for all i, where c is some positive scalar. It follows from (5.5) that $\pi_i(t) \to c\pi_i^*$ for all i. □

Upon obtaining $c\pi^*$, the desired solution π^* can be recovered by normalizing $c\pi^*$.

6. Neural networks. Consider a connected, directed network with node set $\mathcal{N} = \{1, \cdots, n\}$ and arc set $\mathcal{A} \subseteq \mathcal{N} \times \mathcal{N}$. Let us, for each $i \in \mathcal{N}$, denote by $\mathcal{U}(i)$ the set $\{j \mid (j, i) \in \mathcal{A}\}$ of upstream neighbors of i. Let $\sigma_1, \cdots, \sigma_n$ be a set of given scalars and let $\{\lambda_{ij}\}_{(i,j)\in\mathcal{A}}$ be a set of nonzero scalars satisfying $\sum_{j\in\mathcal{U}(i)} |\lambda_{ij}| \leqq 1$ for all i. We wish to find scalars $x_1, \cdots, x_n$ such that

$$x_i = \phi_i\left(\sum_{j\in\mathcal{U}(i)} \lambda_{ij} x_j + \sigma_i\right), \quad \forall i, \tag{6.1}$$

where $\phi_i : \Re \to \Re$ is a continuous nondecreasing function satisfying

$$\lim_{\xi\to-\infty} \phi_i(\xi) = -1, \qquad \lim_{\xi\to+\infty} \phi_i(\xi) = 1, \tag{6.2}$$

(see Fig. 6.1). Notice that the function ϕ_i maps the box $[-1, 1]^n$ into itself and, by Brouwer's fixed point theorem ([11, p. 17]), the system (6.1) is guaranteed to have a solution.

If we think of each node i as a neuron, (6.1) and (6.2) imply that this neuron is turned on (i.e., $x_i \approx 1$) if the majority of its inputs are also turned on. Thus, x_i gives the state ("on" or "off") of the ith neuron for a given set of connections (specified by $\mathcal{A}$) and a given external excitation (specified by σ_i) (see Fig. 6.2.). Indeed, (6.1) and (6.2) describe a class of *neural networks* that have been applied to solving a number of problems in combinatorial optimization, pattern recognition, and artificial intelligence [15]-[16], [19], [25].

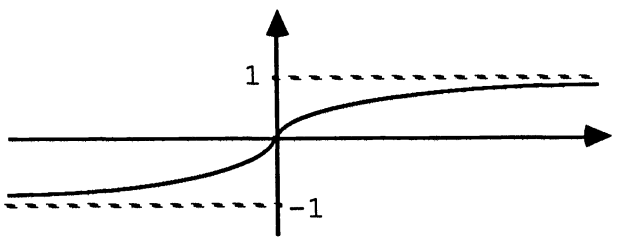


FIG. 6.1. *The function* ϕ_i.

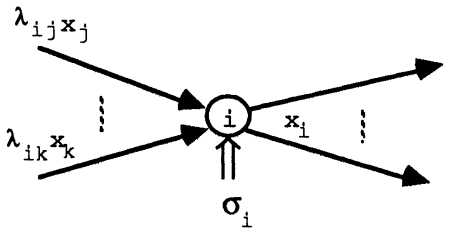


FIG. 6.2

Let $f: \Re^n \to \Re^n$ be the function whose ith component is

$$f_i(x) = \phi_i\left(\sum_{j \in \mathcal{U}(i)} \lambda_{ij} x_j + \sigma_i\right), \quad \forall i. \tag{6.3}$$

Then solving (6.1) is equivalent to finding a fixed point of f. In what follows, we consider a special form for ϕ_i and show that it gives rise, in a natural way, to a nonexpansive function f that satisfies Assumptions B′ and C of § 2. To the best of our knowledge, asynchronous convergence of neural networks has not been explored before. In some sense, asynchronous neural networks are quite natural since biological neural connections may experience long propagation delay [25].

Let ϕ_i^+ denote the *right derivative* of ϕ_i, i.e.,

$$\phi_i^+(\xi) = \lim_{\varepsilon \downarrow 0} (\phi_i(\xi + \varepsilon) - \phi_i(\xi))/\varepsilon, \quad \forall \xi \in \Re.$$

The following result shows that, if ϕ_i^+ is sufficiently small for all i, then f given by (6.3) satisfies Assumption B′.

PROPOSITION 6.1. *If $\mathcal{G}$ is strongly connected and each ϕ_i is continuous, satisfies* (6.2) *and*

$$0 \leqq \phi_i^+(\xi) \leqq 1, \quad \forall \xi \in \Re, \tag{6.4}$$

then f given by (6.3) *satisfies Assumption* B′.

Proof. We have seen earlier that f has a fixed point. Since each ϕ_i is continuous, f is also continuous. Now we will show that f is nonexpansive. Fix any $i \in \mathcal{N}$. Since (cf. (6.4)) the slope of ϕ_i is bounded inside the interval [0, 1], we have

$$|\phi_i(b) - \phi_i(a)| \leqq |b - a|, \quad \forall a \in \Re,\ b \in \Re.$$

Hence, for any $x \in \Re^n$ and $y \in \Re^n$,

$$\begin{aligned} |f_i(y) - f_i(x)| &= \left|\phi_i\left(\sum_{j \in \mathcal{U}(i)} \lambda_{ij} y_j + \sigma_i\right) - \phi_i\left(\sum_{j \in \mathcal{U}(i)} \lambda_{ij} x_j + \sigma_i\right)\right| \\ &\leqq \left|\sum_{j \in \mathcal{U}(i)} \lambda_{ij}(y_j - x_j)\right| \qquad (6.5) \\ &\leqq \sum_{j \in \mathcal{U}(i)} |\lambda_{ij}||y_j - x_j|. \end{aligned}$$

Since $\sum_{j \in \mathcal{U}(i)} |\lambda_{ij}| \leqq 1$, (6.5) implies that

$$|f_i(y) - f_i(x)| \leqq \|x - y\|.$$

Since the choice of i was arbitrary, this in turn implies that

$$\|f(x) - f(y)\| \leqq \|x - y\|, \quad \forall x \in \Re^n,\ y \in \Re^n.$$

Therefore f is nonexpansive.

It remains to show that f satisfies Assumption B'(d). Suppose the contrary. Then for some $x \notin X^*$ and some $x^* \in X^*$, where X^* is the set of fixed points of f, there exists, for every $i \in I(x; x^*)$, an $x^i \in U(x; x^*)$ such that

$$x^i \notin X^* \quad \text{and} \quad f_i(x^i) = x_i^i.$$

Let $J = I(x; x^*)$ ($J \neq \mathcal{N}$ since $x^i \notin X^*$ for all $i \in J$) and $\beta = \|x - x^*\|$. Fix any $i \in J$. By (6.5) and the fact $x^* = f(x^*)$, we obtain that

$$|x_i^i - x_i^*| = |f_i(x^i) - f_i(x^*)| \leqq \sum_{j \in \mathcal{U}(i)} |\lambda_{ij}||x_j^i - x_j^*|.$$

Hence

$$\begin{aligned}\beta &\leqq \sum_{j \in \mathcal{U}(i)} |\lambda_{ij}||x_j^i - x_j^*| \\ &= \sum_{j \in \mathcal{U}(i)} |\lambda_{ij}|\beta + \sum_{j \in \mathcal{U}(i), j \notin J} |\lambda_{ij}|(|x_j^i - x_j^*| - \beta) \\ &\leqq \beta + \sum_{j \in \mathcal{U}(i), j \notin J} |\lambda_{ij}|(|x_j^i - x_j^*| - \beta).\end{aligned}$$

Since $|x_j^i - x_j^*| < \beta$ and $\lambda_{ij} \neq 0$ for all $j \in \mathcal{U}(i)$, $j \notin J$, this implies that $\mathcal{U}(i) \subseteq J$. Since the choice of $i \in J$ was arbitrary, it follows that $\mathcal{U}(i) \subseteq J$ for all $i \in J$. Hence $\mathcal{G}$ is not strongly connected, a contradiction of our hypothesis. □

It follows from Lemmas 2.1, 2.4 and Propositions 2.1, 6.1 that the asynchronous iteration

$$x_i := (1 - \gamma)x_i + \gamma\phi_i\left(\sum_{j \in \mathcal{U}(i)} \lambda_{ij}x_j + \sigma_i\right)$$

(with $0 < \gamma < 1$) converges. Two examples of ϕ_i that satisfy the hypothesis of Proposition 6.1 are

$$\phi_i(\xi) = 2(1 + e^{-2\xi})^{-1} - 1,$$

and

$$\phi_i(\xi) = \max\{-1, \min\{1, \xi\}\}.$$

Let us briefly discuss an alternative form for the function ϕ_i. If we assume that each ϕ_i is continuously differentiable and its derivative $\nabla\phi_i$ satisfies $0 < \nabla\phi_i(\xi) < 1$ for all $\xi \in \Re$, then it can be shown that the restriction of the function f on a compact set is a contraction. In that case, the asynchronous neural iteration

$$x_i := \phi_i\left(\sum_{j \in \mathcal{U}(i)} \lambda_{ij}x_j + \sigma_i\right)$$

can be shown to converge even under the total asynchronism assumption

$$\lim_{t \to +\infty} \tau_{ij}(t) = +\infty, \quad \forall i, \quad \forall j$$

(cf. [7, Chap. 6.2, Prop. 2.1]).

7. Least element of weakly diagonally dominant, Leontief systems. Let $A = [a_{kj}]$ be a given $m \times n$ matrix (with $m \geqq n$) and $b = (b_1, \cdots, b_m)$ be an element of $\Re^m$. We make the following assumption.

Assumption F.

(a) Each row of A has exactly one positive entry and the index set

$$I(i) = \{k \mid a_{ki} > 0\}$$

is nonempty for all i (i.e., every column has at least one positive entry).

(b) $\sum_j a_{kj} \geqq 0$, for all k.

(c) For any $(k_1, \cdots, k_n) \in I(1) \times \cdots \times I(n)$, the $n \times n$ matrix $[a_{k_i j}]$ is irreducible.

Since $a_{ki} > 0$ for all $k \in I(i)$, we will, by dividing the kth constraint by a_{ki} if necessary, assume that $a_{ki} = 1$ for all $k \in I(i)$, in which case parts (a) and (b) of Assumption F are equivalent to

$$a_{ki} = 1, \qquad -\sum_{j \neq i} a_{kj} \leqq 1 \quad \text{and} \quad a_{kj} \leqq 0, \quad \forall j \neq i, \tag{7.1}$$

for all $k \in I(i)$ and all i.

Let X be the polyhedral set

$$X = \{x \in \Re^n \mid Ax \geqq b\}. \tag{7.2}$$

We wish to find an element η of X satisfying

$$x \geqq \eta, \quad \forall x \in X$$

(such an element is called the *least element* of X in [10] and [13]). Notice that if a least element exists, then it is unique. Let $h: \Re^n \to \Re^n$ be the function whose ith component is

$$h_i(x) = \max_{k \in I(i)} \left\{ b_k - \sum_{j \neq i} a_{kj} x_j \right\}. \tag{7.3}$$

It is shown in [10] that X has a least element for all b such that X is nonempty if and only if A^T is *Leontief* (a matrix E is Leontief if each column of E has at most one positive entry and there exists $y \geqq 0$ such that $Ey > 0$ componentwise). The following lemma sharpens this result by giving a necessary and sufficient condition for X to have a least element that is simpler to verify. It also relates the least element of X to the fixed points of h.

LEMMA 7.1. *Suppose that $X \neq \varnothing$ and that Assumption* F *holds. Then,*

(a) *X has no least element if and only if*

$$\sum_j a_{kj} = 0, \quad \forall k. \tag{7.4}$$

(b) *If η is a least element of X, then it is a fixed point of h.*

Proof. We first prove (a). Suppose that (7.4) holds and let $e \in \Re^n$ be the vector with all coordinates equal to 1. Equation (7.4) says that $Ae = 0$. Thus, if x is an element of X, then $x - \lambda e \in X$, for all positive scalars λ. It follows that X cannot have a least element. Now suppose that (7.4) does not hold. We first show that X is bounded from below (i.e., there exists some $a \in \Re^n$ such that $x \geqq a$ componentwise for all $x \in X$). If this were not so, then there would exist some $v \in \Re^n$ and some $x \in X$ such that $v_i < 0$ for some i and $x + \lambda v \in X$ for all positive scalars λ. The latter implies that $A(x + \lambda v) \geqq b$ for all $\lambda > 0$ and hence $Av \geqq 0$. Fix any scalars $(k_1, \cdots, k_n) \in I(1) \times \cdots \times I(n)$ and consider an i such that $v_i = \min_j \{v_j\}$. Then (cf. $Av \geqq 0$)

$$0 \leqq \sum_j a_{k_i j} v_j = \left(\sum_j a_{k_i j} \right) v_i + \sum_{j \neq i} a_{k_i j} (v_j - v_i).$$

Since $v_i < 0$ and $v_j - v_i \geqq 0$ for all $j \neq i$, this, together with the facts (cf. (7.1)) $\sum_j a_{k_i j} \geqq 0$ and $a_{k_i j} \leqq 0$ for all $j \neq i$, implies that $\sum_j a_{k_i j} = 0$ and $v_i = v_j$ for all $j \neq i$ such that $a_{k_i j} \neq 0$. By Assumption F(c), there exists $j \neq i$ such that $a_{k_i j} \neq 0$. We then repeat the above argument with j in place of i. In this way, we eventually obtain that $v_1 = \cdots = v_n$ and $\sum_j a_{k_i j} = 0$ for all i. Since our choice of $(k_1, \cdots, k_n) \in I(1) \times \cdots \times I(n)$ was arbitrary, (7.4) holds—contradicting our hypothesis. Hence X is bounded from below. Using

(7.1), it is easily verified that if x' and x'' are two elements of X, then the n-vector x whose ith component is $\min\{x_i', x_i''\}$ is also an element of X. Since X is closed and bounded from below, X has a least element.

We next prove (b). Since $\eta \in X$, we have (cf. (7.1), (7.2))

$$\sum_{j \neq i} a_{kj}\eta_j + \eta_i \geqq b_k, \quad \forall k \in I(i), \quad \forall i.$$

Thus,

$$h_i(\eta) = \max_{k \in I(i)} \left\{ b_k - \sum_{j \neq i} a_{kj}\eta_j \right\} \leqq \eta_i, \quad \forall i.$$

If η is not a fixed point of h, then the set $I = \{i \mid h_i(\eta) < \eta_i\}$ is nonempty. Then we have

$$\sum_{j} a_{kj}\eta_j > b_k, \quad \forall k \in I(i), \quad \forall i \in I. \tag{7.5}$$

Consider the n-vector $\tilde{\eta}$, defined by $\tilde{\eta}_i = \eta_i - \varepsilon$, if $i \in I$, and $\tilde{\eta}_i = \eta_i$, otherwise. For ε positive and small enough, the inequalities (7.5) remain valid. On the other hand, for all $i \notin I$ and all $k \in I(i)$ we have

$$\sum_{j} a_{kj}\tilde{\eta}_j = \sum_{j \notin I} a_{kj}\eta_j + \sum_{j \in I} a_{kj}(\eta_j - \varepsilon) \geqq \sum_{j} a_{kj}\eta_j \geqq b_k,$$

where we used the property $a_{kj} \leqq 0$ for all j such that $k \notin I(j)$. Thus, $\tilde{\eta} \in X$, contradicting the hypothesis that η is the least element of X. □

Let X^* denote the set of fixed points of h. Suppose that X^* is nonempty (Lemma 7.1 gives sufficient conditions for X^* to be nonempty). We will show that h satisfies Assumption B′. Since (cf. (7.3)) h is continuous, it suffices to show that parts (c) and (d) of Assumption B′ hold.

LEMMA 7.2. *$\|h(x) - h(y)\| \leqq \|x - y\|$ for any $x \in \Re^n$ and any $y \in \Re^n$.*

Proof. Let $z = h(x)$, $w = h(y)$ and consider any $i \in \{1, \cdots, n\}$. We will show that $|z_i - w_i| \leqq \|x - y\|$, from which our claim follows. Since $z_i = h_i(x)$ and $w_i = h_i(y)$, it follows from (7.3) that, for some k in $I(i)$,

$$\sum_{j \neq i} a_{kj}x_j + z_i \geqq b_k, \tag{7.6a}$$

$$\sum_{j \neq i} a_{kj}y_j + w_i = b_k. \tag{7.6b}$$

Subtracting (7.6b) from (7.6a), we obtain

$$\sum_{j \neq i} a_{kj}(x_j - y_j) + (z_i - w_i) \geqq 0.$$

This together with (7.1) implies that

$$\begin{aligned} w_i - z_i &\leqq \sum_{j \neq i} a_{kj}(x_j - y_j) \\ &\leqq \sum_{j \neq i} |a_{kj}| \, \|x - y\| \\ &\leqq \|x - y\|. \end{aligned}$$

The inequality $z_i - w_i \leqq \|x - y\|$ is obtained similarly. □

LEMMA 7.3. *h satisfies Assumption* B′(d).

Proof. Suppose the contrary. Then for some $x \notin X^*$ and some $x^* \in X^*$, there exists, for every $i \in I(x; x^*)$, an $x^i \in U(x; x^*)$ such that

$$x^i \notin X^* \quad \text{and} \quad h_i(x^i) = x_i^i.$$

Let $J = I(x; x^*)$, $J^- = \{i \mid x_i - x_i^* = -\beta\}$, $J^+ = \{i \mid x_i - x_i^* = \beta\}$ and $\beta = \|x - x^*\|$. (We must have $J \neq \{1, \cdots, n\}$ because otherwise the set $U(x; x^*)$ would be a singleton, implying that the vectors $x^1, \cdots, x^n$ are all equal, in which case each x^i is a fixed point of h, a contradiction.)

Fix any $i \in J^-$. By (7.3) and the hypothesis $x^* = h(x^*)$, there exists some $k_i \in I(i)$ such that

$$\sum_j a_{k_i j} x_j^* = b_{k_i}.$$

Since $x_i^i = h_i(x^i)$, we then have $\sum_j a_{k_i j} x_j^i \geqq b_{k_i} = \sum_j a_{k_i j} x_j^*$, so

$$\sum_j a_{k_i j}(x_j^i - x_j^*) \geqq 0.$$

This implies (using (7.1) and the facts $k_i \in I(i)$, $i \in J^-$) that

$$\begin{aligned} 0 &\leqq -\beta \sum_{j \in J^-} a_{k_i j} + \beta \sum_{j \in J^+} a_{k_i j} + \sum_{j \notin J} |a_{k_i j}||x_j^i - x_j^*| \\ &= -\beta \sum_{j \in J^-} a_{k_i j} - \beta \sum_{j \in J^+} |a_{k_i j}| + \sum_{j \notin J} |a_{k_i j}||x_j^i - x_j^*| \\ &= -\beta \left(1 - \sum_{j \neq i} |a_{k_i j}|\right) - 2\beta \sum_{j \in J^+} |a_{k_i j}| + \sum_{j \notin J} |a_{k_i j}|(|x_j^i - x_j^*| - \beta). \end{aligned}$$

Since $|x_j^i - x_j^*| < \beta$ for all $j \notin J$, (7.1) implies that

$$\sum_{j \neq i} a_{k_i j} = -1 \quad \text{and} \quad a_{k_i j} = 0, \quad \forall j \notin J^-. \tag{7.7}$$

Since the choice of i was arbitrary, (7.7) holds for all $i \in J^-$. By an analogous argument, we also obtain that, for all $i \in J^+$,

$$\sum_{j \neq i} a_{k_i j} = -1 \quad \text{and} \quad a_{k_i j} = 0, \quad \forall j \notin J^+, \tag{7.8}$$

where each k_i is a scalar in $I(i)$ such that

$$\sum_j a_{k_i j} x_j^i = b_{k_i}.$$

For each $i \notin J$, let k_i be any element of $I(i)$. Since $J \neq \{1, \cdots, n\}$, (7.7) and (7.8) imply that the $n \times n$ matrix $[a_{k_i j}]_{i,j}$ is *not* irreducible—a contradiction of Assumption F(c). □

We may now invoke Lemmas 2.1, 2.4 and Proposition 2.1 to establish that the partially asynchronous iteration

$$x := (1 - \gamma)x + \gamma h(x)$$

(with $0 < \gamma < 1$) converges to a fixed point of h. Unfortunately, such a fixed point is not necessarily the least element of X. We have, however, the following characterization of such fixed points.

LEMMA 7.4. *If X has a least element η, then, for any fixed point x^* of h, there exists a nonnegative scalar λ such that $x^* = \eta + (\lambda, \cdots, \lambda)$.*

Proof. Since x^* is a fixed point of h, $x^* \in X$. Hence $x^* \geqq \eta$. We then repeat the proof of Lemma 7.3, with $J^- = \{1, \cdots, n\}$ and $x^i = \eta$ for all i. This yields that, for every $i \in \{1, \cdots, n\}$, there exists some $k_i \in I(i)$ such that $x_i^* - \eta_i \leqq \sum_{j \neq i} |a_{k_i j}|(x_j^* - \eta_j)$. Since $x^* - \eta \geqq 0$, Assumption F(c) and (7.1) imply that the $x_i^* - \eta_i$'s are equal. □

Lemma 7.4 states that, given a fixed point x^* of h, we can compute the least element of X by a simple line search along the direction $(-1, \cdots, -1)$ (the stepsize λ is the largest for which $x^* - (\lambda, \cdots, \lambda)$ is in X). An example of X for which the corresponding h has multiple fixed points is

$$X = \{(x_1, x_2) \mid x_1 - x_2 \geqq 0,\ x_1 - 0.5x_2 \geqq -1,\ -x_1 + x_2 \geqq 0\}.$$

Here $h_1(x) = \max\{x_2, 0.5x_2 - 1\}$, $h_2(x) = x_1$ and both $(-1, -1)$ and $(-2, -2)$ are fixed points of h (the least element of X is $(-2, -2)$).

Let us remark that if the inequalities in Assumption F(b) are strict, then the mapping h is a contraction mapping (the same argument as in Lemma 7.2) and convergence under total asynchronism is obtained. We also remark that, if in the statement of Assumption F(c) we replace "For any" by the weaker "For some," then Lemmas 7.1 and 7.2 still hold, but Lemmas 7.3 and 7.4 do not. In fact, it can be shown that X^* is not necessarily convex in this case.

8. Simulation for network flow problems. In this section we study and compare, using simulation, the performance of synchronous and partially asynchronous algorithms for the network flow problem of § 4. We measure the following: (a) the effects of the stepsize γ (cf. Lemma 2.1), the problem size n, and the asynchrony measure B on the performance of partially asynchronous algorithms, (b) the efficiency of different partially asynchronous algorithms relative to each other and also relative to the corresponding synchronous algorithms.

In our study, we consider a special case of the network flow problem (4.1)-(4.2) where each cost function $a_{ij}(\cdot)$ is a quadratic on $[0, +\infty]$, i.e.,

$$a_{ij}(f_{ij}) = \begin{cases} \alpha_{ij}|f_{ij}|^2 + \beta_{ij} f_{ij} & \text{if } f_{ij} \geqq 0, \\ +\infty & \text{otherwise,} \end{cases} \tag{8.1}$$

where α_{ij} is a given positive scalar and β_{ij} is a given scalar. This special case has many practical applications and has been studied extensively [6], [9], [12], [21], [31]. In what follows, we will denote by $h: \Re^n \to \Re^n$ the function given by (4.3), (4.5)-(4.6), and (8.1). All of the algorithms involved in our study are based on h.

8.1. Test problem generation. In our test, each α_{ij} is randomly generated from the interval $[1, 5]$ and each β_{ij} is randomly generated from the set $\{1, 2, \cdots, 100\}$. The number of arcs is ten times the number of nodes and the average node supply is 1000, i.e., $|s_1| + \cdots + |s_n| = 1000n$. Half of the nodes are supply nodes and half of the nodes are demand nodes (we say a node i is a supply (demand) node if $s_i > 0$ ($s_i < 0$)). The problems are generated using the linear cost network generator NETGEN [18], modified to generate quadratic cost coefficients as well.

8.2. The main partially asynchronous algorithm. The main focus of our study is the partially asynchronous algorithm described in § 4. This algorithm, called PASYN, generates a sequence $\{x(t)\}$ using the partially asynchronous iteration (1.1) under Assumption A, where the algorithmic mapping f is given by

$$f(x) = (1 - \gamma)x + \gamma h(x). \tag{8.2}$$

In our simulation, the communication delays $t-\tau_{ij}(t)$ are independently generated from a uniform distribution on the set $\{0, 1, \cdots, B-1\}$ and, for simplicity, we assume that $\mathcal{T}_i = \{1, 2, \cdots\}$ for all i. (This models a situation where the computation delay is negligible compared to the communication delay.) The components of $x(1-B)$, $x(2-B), \cdots, x(0)$ are independently generated from a uniform distribution over the interval $[0, 10]$ (this is to reflect a lack of coordination among processors) and the algorithm terminates at time t if $\max_{\tau,\tau' \in \{t-B,\cdots,t\}} \|x(\tau)-x(\tau')\| \leqq 0.001$.

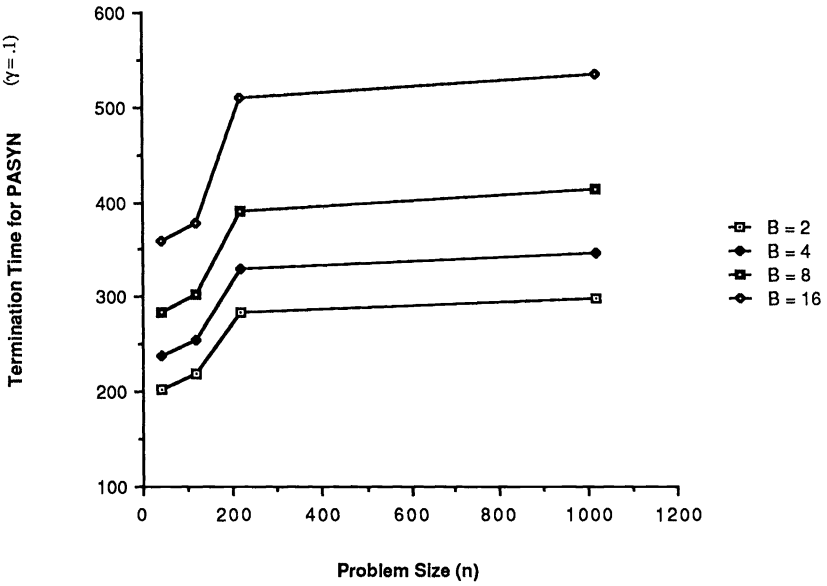


FIG. 8.1(a). *Termination time for* PASYN ($\gamma = 0.1$), *for different values of B and n.*

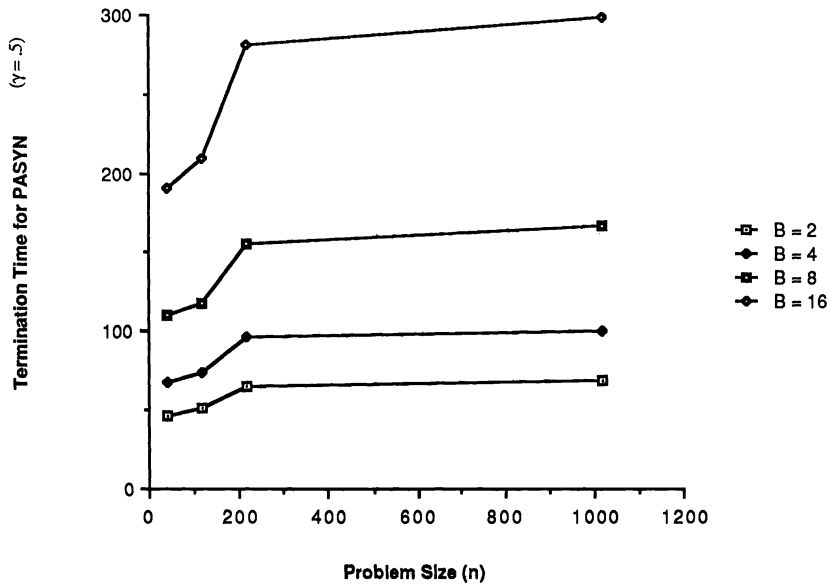


FIG. 8.1(b). *Termination time for* PASYN ($\gamma = 0.5$), *for different values of B and n.*

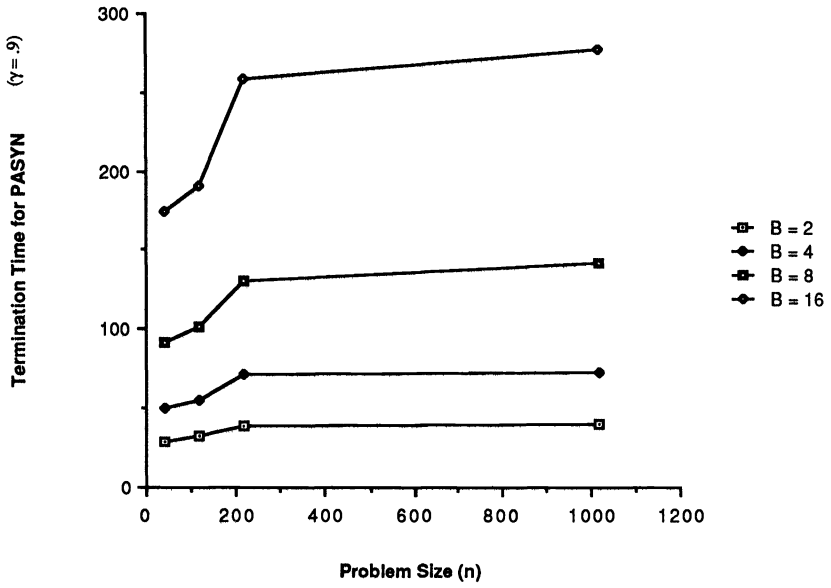


FIG. 8.1(c). *Termination time for* PASYN ($\gamma = 0.9$), *for different values of B and n.*

The termination time of PASYN, for different values of γ, B, and n, is shown in Figs. 8.1(a)-(c). In general, the rate of convergence of PASYN is the fastest for γ near 1 and for B small, corroborating our intuition. The termination time grows quite slowly with the size of the problem n but quite fast with decreasing γ. For γ near 1, the termination time grows roughly linearly with B (but not when γ is near 0).

8.3. An alternative partially asynchronous algorithm. Consider the function $f^0: \Re^n \to \Re^n$ whose ith component is given by

$$f_i^0(x) = \begin{cases} h_i(x) & \text{if } i \neq 1, \\ x_1 & \text{otherwise.} \end{cases} \tag{8.3}$$

It is shown in [5] that the algorithm $x := f^0(x)$ converges under the total asynchronism assumption. Hence it is of interest to compare this algorithm with that described in § 8.2 (namely PASYN) under the same assumption of partial asynchronism. The partially asynchronous version of the algorithm $x := f^0(x)$, called TASYN, is identical to PASYN except that the function f in (8.2) is replaced by f^0. (Note that TASYN has the advantage that it uses a unity stepsize.)

The termination time of TASYN, for different values of B and n, is shown in Fig. 8.2. A comparison with Figs. 8.1(a)-(c) shows that TASYN is considerably slower than PASYN. The speed of TASYN is improved if f in (8.2) is replaced by f^0 only after a certain amount of time has elapsed, but the improvement is still not sufficient for it to compete with PASYN.

8.4. Two synchronous algorithms. In this subsection we consider two types of synchronous algorithms based on h: the Jacobi algorithm and the Gauss-Seidel algorithm. In particular, the Gauss-Seidel algorithm has been shown to be efficient for practical computation (see [6], [9], [21], [31]). Hence, by comparing the asynchronous algorithms with these algorithms, we can better measure the practical efficiency of the former.

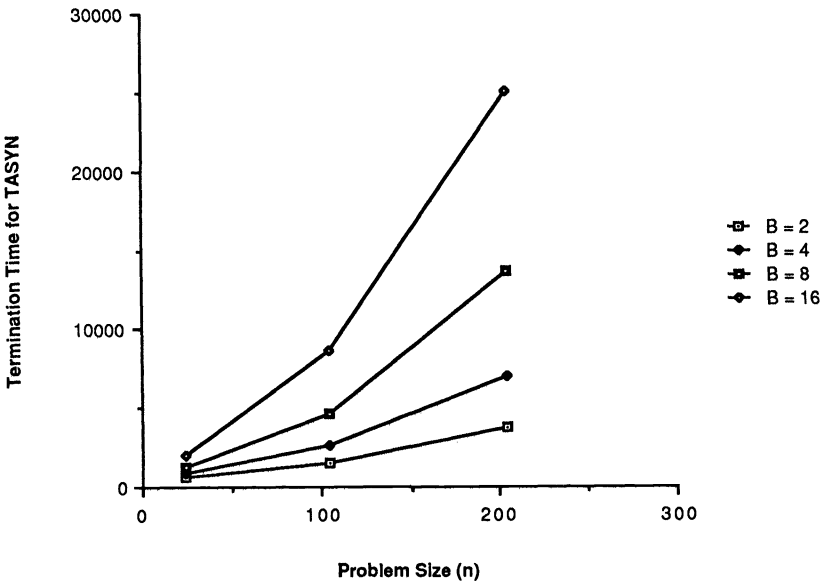


FIG. 8.2. *Termination time for* TASYN, *for different values of B and n.*

The Jacobi algorithm, called SYNJB, is a parallel algorithm that generates a sequence $\{x(t)\}$ according to

$$x(t+1) = (1-\gamma)x(t) + \gamma h(x(t)),$$

where $\gamma \in (0, 1)$. The initial estimates $x_1(0), \cdots, x_n(0)$ are independently generated from a uniform distribution over the interval [0, 10], and the algorithm terminates at time t if $\|x(t) - x(t-1)\| \leqq 0.001$. (SYNJB can be seen to be a special case of PASYNB where $B = 1$ and hence $\{x(t)\}$ converges to a fixed point of h.)

Consider any positive integer b and any function $\beta : \{1, \cdots, n\} \to \{1, \cdots, b\}$ such that $h_i(x)$ does not depend on x_j if $\beta(i) = \beta(j)$. We associate with b and β a Gauss-Seidel algorithm that generates a sequence $\{x(t)\}$ according to

$$x_i(t+1) = \begin{cases} h_i(x(t)) & \text{if } t \equiv \beta(i) - 1 \pmod{b}, \\ x_i(t) & \text{otherwise.} \end{cases}$$

In our simulation, the initial estimates $x_1(0), \cdots, x_n(0)$ are independently generated from a uniform distribution over the interval [0, 10] and the algorithm terminates at time t if

$$\max_{\tau, \tau' \in \{t-b, \cdots, t\}} \|x(\tau) - x(\tau')\| \leqq 0.001.$$

(Convergence of $\{x(t)\}$ to a fixed point of h follows from Proposition 2.4 in [6]. Note that, similar to TASYN, this algorithm has the advantage of using a unity stepsize.) We consider both a serial and a parallel version of this algorithm (this is done by choosing b and β appropriately). SYNGS1 is the serial version which chooses $b = n$ and $\beta(i) = i$ for all i. SYNGS2 is the parallel version which uses a coloring heuristic to find, for each problem, a choice of b and β for which b is small.

The termination time for SYNJB, SYNGS1 and SYNGS2, for different values of n, are shown in Figs. 8.3(a)-(b). In Fig. 8.3(a), the choice of b obtained by the coloring heuristic in SYNGS2 is also shown (in parentheses). In general, SYNJB is considerably

faster than either of the two Gauss-Seidel algorithms SYNGS1 and SYNG2 (however in SYNJB all processors must compute at all times). From Fig. 8.3(b) we see that, as n increases and the problems become more sparse, SYNGS2 (owing to its high parallelism) becomes much faster than the serial algorithm SYNGS1. (Notice that the time for SYNGS1 is approximated by the time for SYNGS2 multiplied by n/b, as expected.) Comparing Fig. 8.3(a) with Fig. 8.1(c), we see that SYNJB is approximately

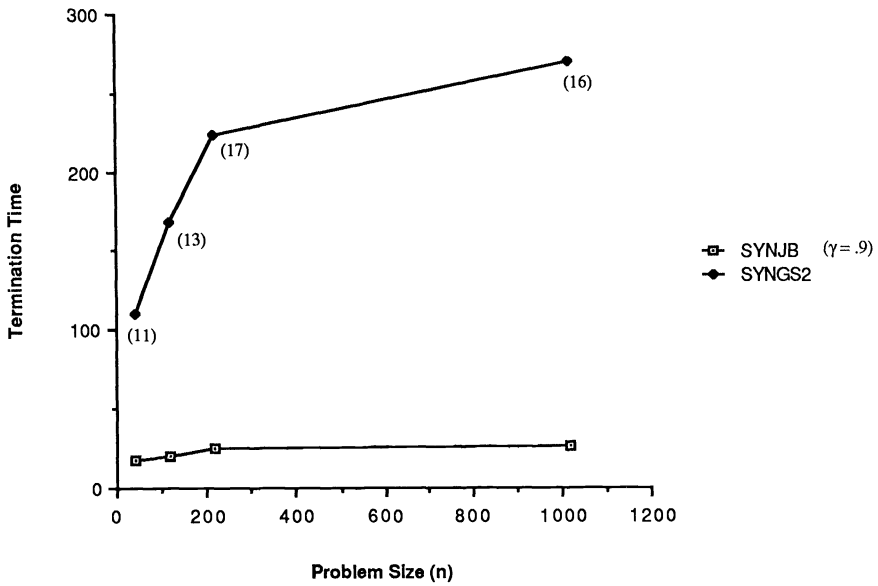


FIG. 8.3(a). *Comparing the termination time for the two synchronous, parallel algorithms* SYNJB ($\gamma = 0.9$) *and* SYNGS2, *for different values of n.*

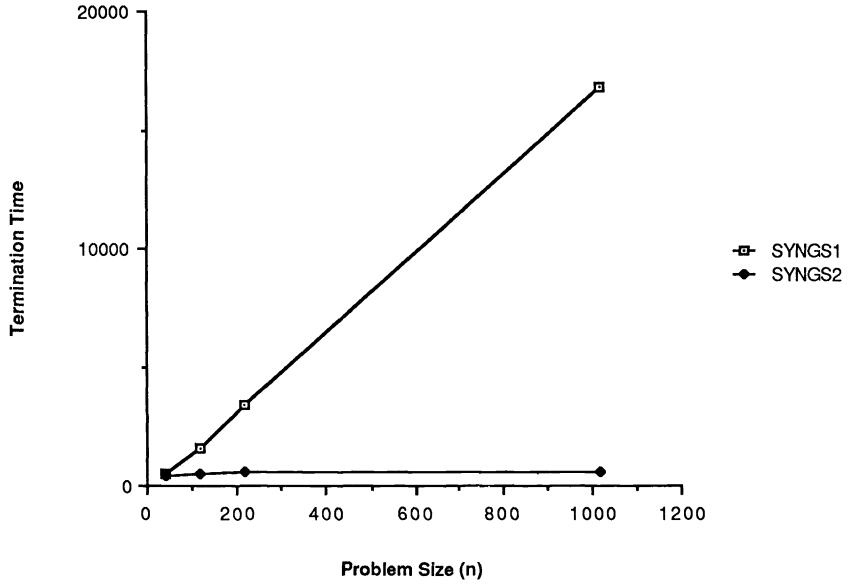


FIG. 8.3(b). *Comparing the termination time for the serial algorithm* SYNGS1 *and for the synchronous, parallel algorithm* SYNGS2, *for different values of n.*

3/2 times faster than PASYN and that PASYN is faster than SYNGS2, unless PASYN suffers long delays.

8.5. Simulation of synchronous algorithms in the face of communication delays. In this subsection we consider the execution of the synchronous iterations of § 8.4 in an asynchronous computing environment, that is, in an environment where communication delays are variable and unpredictable. The mathematical description of the algorithms in this subsection is identical to that of the algorithms considered in the preceding subsection; for this reason, the number of iterations until termination is also the same. On the other hand, each processor must wait until it receives the updates of the other processors before it can proceed to the next iteration. For this reason, the actual time until termination is different from the number of iterations. In our simulation, the delays are randomly generated but their statistics are the same as in our simulation of asynchronous algorithms in §§ 8.2 and 8.3 (uniformly distributed over the set $\{0, 1, \cdots, B-1\}$, where B denotes the maximum delay). This will allow us to determine whether asynchronous methods are preferable in the face of communication delays.

More precisely, consider any synchronous algorithm and let T denote the number of iterations at which this algorithm terminates. With each $t \in \{1, \cdots, T\}$ and each $i \in \{1, \cdots, n\}$, we associate a positive integer $\sigma_i(t)$ to represent the "time" at which the update of the ith component at iteration t is performed in the corresponding asynchronous execution. (Here we distinguish between "iteration" for the synchronous algorithm and "time" for the asynchronous execution.) Then $\{\sigma_i(t)\}$ is recursively defined by the following formula:

$$\sigma_i(t) = \max \{\sigma_j(t-1) + (\text{communication delay from proc. } j \text{ to proc. } i \text{ at time } \sigma_j(t-1))\},$$

where the maximization is taken over all j such that the jth component influences the ith component at iteration t. The termination time of the asynchronous algorithm is then taken to be

$$\max_i \{\sigma_i(T)\}.$$

The partially asynchronous algorithms that simulate SYNJB, SYNGS1 and SYNGS2 are called, respectively, PASYNJB, PASYNGS1 and PASYNGS2. The termination times for these algorithms are shown in Figs. 8.4–8.6 (they are obtained from the termination times shown in Figs. 8.3(a)–(b) using the procedure described above). Comparing these figures with Figs. 8.1(a)–(c), we see that PASYNJB is roughly 3/4 as fast as PASYN (when both use the same stepsize $\gamma = 0.9$) while the other two algorithms PASYNGS1 and PASYNGS2 are considerably slower than PASYN (even when PASYN uses the most conservative stepsize $\gamma = 0.1$).

To summarize, we can conclude that PASYN is the fastest algorithm for partially asynchronous computation and that its synchronous counterpart SYNJB is the fastest for synchronous parallel computation. We remark that similar behavior was observed in other network flow problems that were generated. Furthermore, the asynchronous algorithm PASYN seems to be preferable to its synchronous counterpart SYNJB in the face of delays. In practice, the assumption that the delays are independent and identically distributed might be violated. For example, queueing delays are usually dependent; also, the distance between a pair of processors who need to communicate could be variable, in which case the delays are not identically distributed. On the other hand, such aspects cannot be simulated convincingly without having a particular parallel computing system in mind.

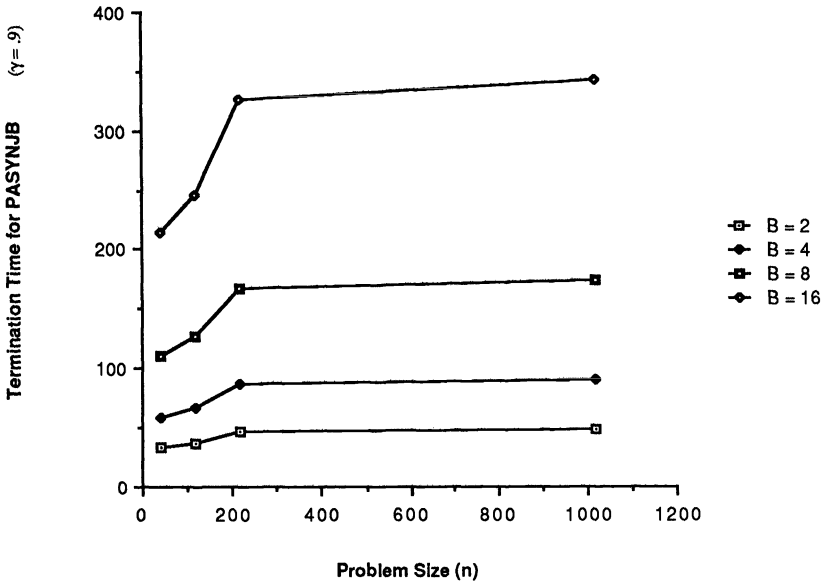


FIG. 8.4. *Termination time for* PASYNJB ($\gamma = 0.9$), *for different values of B and n.*

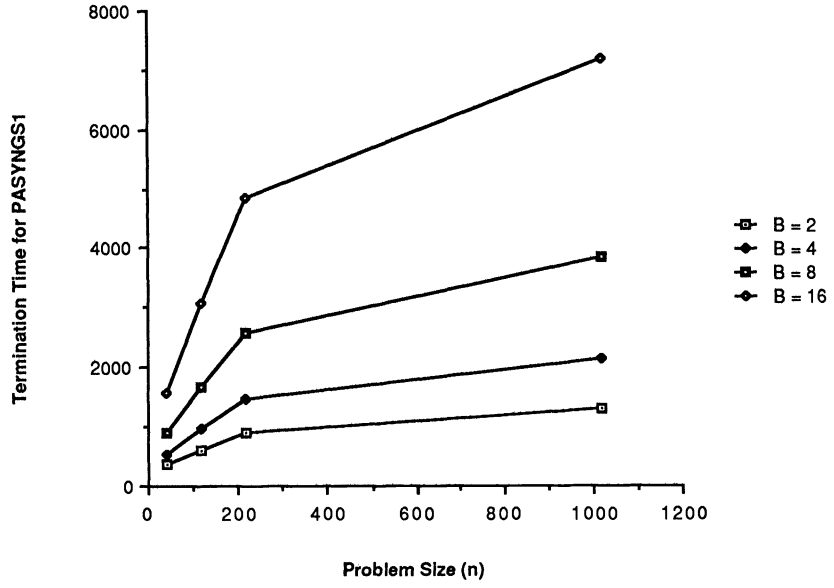


FIG. 8.5. *Termination time for* PASYNGS1, *for different values of B and n.*

9. Conclusion and extensions. In this paper we have presented a general framework, based on nonexpansive mappings, for partially asynchronous computation. The key to this framework is a new class of functions that are nonexpansive with respect to the maximum norm. We showed that any algorithm whose algorithmic mapping belongs to this class converges under the partial asynchronism assumption with an arbitrarily large bound on the delays. While some of the asynchronous algorithms thus obtained are known, others are quite new. Numerical experimentation

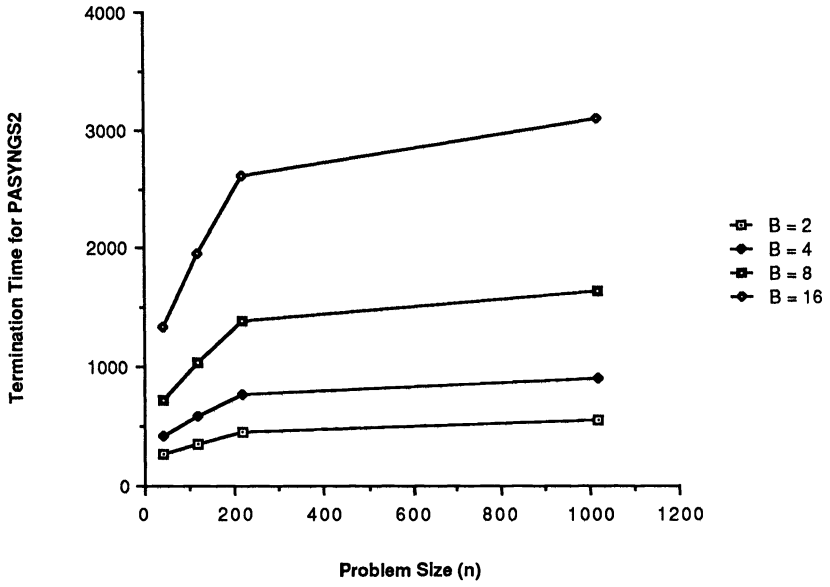


FIG. 8.6. *Termination time for* PASYNGS2, *for different values of B and n.*

with network flow problems suggests that, for partially asynchronous computation, the new algorithms may be substantially faster than those obtained from synchronous algorithms.

REFERENCES

[1] G. M. BAUDET, *Asynchronous iterative methods for multiprocessors*, J. Assoc. Comput. Mach., 15 (1978), pp. 226-244.

[2] D. P. BERTSEKAS, *Distributed dynamic programming*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 610-616.

[3] ———, *Distributed asynchronous computation of fixed points*, Math. Programming, 27 (1983), pp. 107-120.

[4] ———, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[5] D. P. BERTSEKAS AND D. EL BAZ, *Distributed asynchronous relaxation methods for convex network flow problems*, SIAM J. Control Optim., 25 (1987), pp. 74-85.

[6] D. P. BERTSEKAS, P. HOSEIN, AND P. TSENG, *Relaxation methods for network flow problems with convex arc costs*, SIAM J. Control Optim., 25 (1987), pp. 1219-1243.

[7] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[8] D. CHAZAN AND W. MIRANKER, *Chaotic relaxation*, Linear Algebra Appl., 2 (1969), pp. 199-222.

[9] R. W. COTTLE, S. G. DUVALL, AND K. ZIKAN, *A Lagrangean relaxation algorithm for the constrained matrix problem*, Naval Res. Logist. Quart., 33 (1986), pp. 55-76.

[10] R. W. COTTLE AND A. F. VEINOTT, JR., *Polyhedral sets having a least element*, Math. Programming, 3 (1972), pp. 238-249.

[11] K. DEIMLING, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, New York, 1985.

[12] R. S. DEMBO AND J. G. KLINCEWICZ, *A scaled reduced gradient algorithm for network flow problems with convex separable costs*, Math. Programming Stud., 15 (1981), pp. 125-147.

[13] H. GABBAY, *A note on polyhedral sets having a least element*, Math. Programming, 11 (1976), pp. 94-96.

[14] F. R. GANTMACHER, *The Theory of Matrices*, Vol. II, Chelsea, New York, 1960.

[15] J. J. HOPFIELD, *Neurons with graded response have collective computational properties like those of two-state neurons*, Proc. Nat. Acad. Sci. U.S.A., 81 (1984), pp. 3088-3092.

[16] J. J. HOPFIELD AND D. W. TANK, *Computing with neural circuits: a model*, Science, 233 (1986), pp. 625-633.

[17] K. R. JAMES, *Convergence of matrix iterations subject to diagonal dominance*, SIAM J. Numer. Anal., 10 (1973), pp. 478-484.

[18] D. KLINGMAN, A. NAPIER, AND J. STUTZ, NETGEN—*A program for generating large scale (un)capacitated assignment, transportation and minimum cost flow network problems*, Management Sci., 20 (1974), pp. 814-822.

[19] R. P. LIPPMANN, *An introduction to computing with neural nets*, IEEE ASSP Magazine, (1987), pp. 4-22.

[20] B. LUBACHEVSKY AND D. MITRA, *A chaotic, asynchronous algorithm for computing the fixed point of a nonnegative matrix of unit spectral radius*, J. Assoc. Comput. Mach., 33 (1986), pp. 130-150.

[21] A. OHUCHI AND I. KAJI, *Lagrangian dual coordinatewise maximization algorithm for network transportation problems with quadratic costs*, Networks, 14 (1984), pp. 515-530.

[22] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[23] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[24] ———, *Network Flows and Monotropic Optimization*, Wiley-Interscience, New York, 1984.

[25] T. J. SEJNOWSKI, *Open questions about computation in cerebral cortex*, in Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. II, J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, eds., MIT Press, Cambridge, MA, 1986, pp. 372-389.

[26] P. TSENG, *Distributed computation for linear programming problems satisfying a certain diagonal dominance condition*, Working Paper 1256, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, Canada; also LIDS Report, MIT, Cambridge, MA, December 1986; Math. Oper. Res., to appear.

[27] J. N. TSITSIKLIS AND D. P. BERTSEKAS, *Distributed asynchronous optimal routing in data networks*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 325-332.

[28] J. N. TSITSIKLIS, D. P. BERTSEKAS, AND M. ATHANS, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 803-812.

[29] J. N. TSITSIKLIS, *Problems in decentralized decision making and computation*, Ph.D. Thesis, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1984.

[30] ———, *On the stability of asynchronous iterative processes*, Mathematical Systems Theory, 20 (1987), pp. 137-153.

[31] S. A. ZENIOS AND J. M. MULVEY, *Relaxation techniques for strictly convex network problems*, Ann. Oper. Res. 5: Algorithms and Software for Optimization C. L. Monma, ed., Baltzer, Basel, Switzerland, 1986, pp. 517-538.

[32] S. A. ZENIOS AND J. M. MULVEY, *A distributed algorithm for convex network optimization problems*, Parallel Comput., 6 (1988), pp. 45-56.

[33] S. A. ZENIOS AND R. A. LASKEN, *Nonlinear network optimization on a massively parallel connection machine*, in Ann. Oper. Res. 14: Parallel Optimization on Novel Computer Architectures, R. R. Meyer and S. A. Zenios, eds., Baltzer, Basel, Switzerland (1988), pp. 147-165.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 711-719, May 1990

013

ON THE EXISTENCE OF LINEAR OPTIMAL CONTROL WITH OUTPUT FEEDBACK*

GUOXIANG GU†

Abstract. A necessary and sufficient condition is established for the existence of an output feedback $u = -Ky$ that is LQ optimal with respect to some nonsingular quadratic cost, a fact that entails robustness in the closed-loop systems. It is shown that if the system (with equal number of inputs and outputs) is strict minimum phase and has nonsingular high-frequency gain, then there exists a linear static output feedback that minimizes a certain quadratic cost functional. The minimization of H_∞ norm of the closed-loop system with output feedback is also discussed.

Key words. optimal control, output feedback, minimum phase systems, high-frequency gain, positive real functions

AMS(MOS) subject classifications. 49, 93

1. Introduction.

Consider the linear system

$$\dot{x} = Ax + Bu, \qquad y = Cx \tag{1.1}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the control, and $y \in \mathbb{R}^p$ is the measured output with $p = m$. Without loss of generality, it is assumed that the input matrix B has full column rank and the realization $\{A, B, C\}$ is both controllable and observable (this can be replaced by stabilizability and detectability). The main purpose of this paper is to investigate the existence of an output feedback control law $u = -Ky$, $K \in \mathbb{R}^{p\times p}$, which not only stabilizes the closed-loop system, but also minimizes a quadratic cost functional

$$J(u) = \int_0^\infty x^T Qx + u^T Ru\, dt, \tag{1.2}$$

for some positive-definite $R \in \mathbb{R}^{p\times p}$ and nonnegative $Q \in \mathbb{R}^{n\times n}$. An important reason to study such an optimal output feedback regulator problem is due to the robustness consideration. It has been shown by Kalman [5] for single-input systems that if a linear state feedback control law $u = -Fx$ minimizes (1.2) for some Q and R, then the system admits an infinite gain margin and a sixty-degree phase margin. An extension of this result to multi-input systems is the Kalman–Anderson inequality [1]. This optimality property can tolerate some system uncertainties such as parameter variations, etc. Therefore recovering this robustness property in absence of states measurement has received great attention [2], [6].

In this paper, a necessary and sufficient condition will be established for the existence of optimal feedback control. The extension of this result to H_∞ control via a linear, static output feedback will be studied in terms of H_∞ sub-optimal control. An example will be given for illustration.

2. Existence of optimal output feedback control.

Let the linear system be given as in (1.1) and the matrices Q and R be defined as in (1.2). Suppose $\{Q, A\}$ is observable. It is well known that the state feedback law $u = -Fx$ is optimal (which minimizes (1.2)), if and only if

$$F = R^{-1}B^T P, \tag{2.1}$$

* Received by the editors February 21, 1989; accepted for publication (in revised form) July 12, 1989.
† Department of Electrical Engineering, Wright State University, Dayton, Ohio 45435.

where $P>0$ is the unique stabilizing solution of the following algebraic Riccati equation:

$$A^TP+PA-PBR^{-1}B^TP+Q=0. \tag{2.2}$$

If an output feedback control law $u=-Ky$ is used, then (2.1) is reduced to

$$F=KC=R^{-1}B^TP. \tag{2.3}$$

The next result is obvious.

LEMMA 2.4. *Let the open-loop system be defined as in* (1.1). *There exists a* $p\times p$ *real matrix* K *such that* $u=-Ky$ *minimizes a certain quadratic cost functional* $J(u)$ *in* (1.2), *if and only if the closed-loop system is internally stabilized and the equations* (2.2) *and* (2.3) *are satisfied for some* $n\times n$ *matrix* $P>0$, $p\times p$ *matrix* $R>0$, *and* $n\times n$ *matrix* $Q\geqq 0$ *with* $\{Q, A\}$ *observable.*

Clearly, the condition in the above result is not satisfactory. The real objective should be an equivalent condition imposed on realization $\{A, B, C\}$ instead of others. The following notion is important.

DEFINITION 2.5. A square transfer function matrix $T(s)=C(sI-A)^{-1}B$ is called positive real if all the eigenvalues of A are on the open left half plane and

$$T^T(-j\omega)+T(j\omega)\geqq 0 \quad \forall\omega\in\mathbb{R}. \tag{2.6}$$

The transfer function matrix $T(s)$ is called strictly positive real if all the eigenvalues of A are on the open left half plane and

$$T^T(-j\omega)+T(j\omega)>0 \quad \forall\omega\in\mathbb{R}. \tag{2.7}$$

The next result characterizes linear quadratic optimal control in terms of positive realness of the closed-loop system [1].

LEMMA 2.8. *A state feedback law* $u=-Fx$ *minimizes* $J(u)$ *in* (1.2) *for some* $R>0$, *and* $Q\geqq 0$ *with* $\{Q, A\}$ *observable, if and only if the system transfer function matrix* $F(sI-A+BF)^{-1}B$ *is positive real. That is, there exists a positive-definite matrix* P *such that* [1]

$$(A-BF)^TP+P(A-BF)=\Phi \quad \textit{and} \quad F=B^TP, \tag{2.9}$$

with Φ *nonpositive and* $\{-\Phi, A\}$ *observable. If* (2.9) *is true, then* $u=-Fx$ *minimizes the cost functional* (1.2) *for* $Q=-\Phi$ *and* $R=I_p/2$.

Clearly, the existence of optimal output feedback is equivalent to the existence of the matrix K such that the transfer function $KC(sI-A+BKC)^{-1}B$ is positive real. In light of the above lemma, the following result is obtained.

THEOREM 2.10. *Let the system be defined as in* (1.1), *which is absent of transmission zeros on an imaginary axis. There exists a matrix* $K\in\mathbb{R}^{p\times p}$ *such that* $u=-Ky$ *minimizes* $J(u)$ *in* (1.2), *for some* $Q\geqq 0$, *and* $R>0$ *with* $\{Q, A\}$ *observable, if and only if* $\det(CB)\neq 0$, *and the open-loop transfer function* $C(sI-A)^{-1}B$ *is minimum phase.*

Proof. (*Sufficiency.*) Without loss of generality, we assume that the input matrix $B^T=[I_p\ \ 0]$, because $\det(CB)\neq 0$. (Otherwise, it is always possible to find a similarity transform to make it true.) With the conditions given, we prove that there exists a positive-definite matrix S, a nonsingular matrix $\mathcal{K}$, and a scalar $\rho>0$, such that

$$\text{(i)}\quad S(A-\rho B\mathcal{K}C)^T+(A-\rho B\mathcal{K}C)S<0, \quad \text{and} \tag{2.11}$$

$$\text{(ii)}\quad CS=\mathcal{K}^{-1}B^T,$$

for which the output feedback gain $K = \rho\mathcal{K}$. It is noted that with $\mathcal{K}^{-1}$ defined in (ii), $\mathcal{K}^{-1}$ is nonsingular. Clearly, if (2.11) is true, then the transfer function $\mathcal{K}C(sI - A + \rho B\mathcal{K}C)^{-1}B$ is strictly positive real, and hence $u = -\rho\mathcal{K}y = -Ky$ is optimal for some $R > 0$, and $Q > 0$, which implies $\{Q, A\}$ observable (see Lemma 2.8 with $S = P^{-1}$, $F = \rho\mathcal{K}C$, and $Q = -\Phi$). Define

$$\Phi := S(A - \rho B\mathcal{K}C)^T + (A - \rho B\mathcal{K}C)S = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}, \tag{2.12}$$

where Φ is partitioned in a way that $\Phi_{11} \in \mathbb{R}^{p\times p}$. Clearly, Φ being negative definite is equivalent to

$$\Phi_{11} < 0, \quad \Phi_{22} < 0, \quad \Phi_{11} - \Phi_{12}\Phi_{22}^{-1}\Phi_{21} < 0. \tag{2.13}$$

Let

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

which has the same partition as matrix Φ. Then from (ii) of (2.11), we obtain

$$C_1S_{11} + C_2S_{21} = \mathcal{K}^{-1} \quad \text{and} \quad C_1S_{12} + C_2S_{22} = 0, \tag{2.14}$$

where $C = [C_1 \;\; C_2]$ with $C_1 \in \mathbb{R}^{p\times p}$ nonsingular (recall that $\det(CB) \neq 0$). By substitution of $S_{12} = -C_1^{-1}C_2S_{22}$, we have

$$\Phi_{22} = (A_{22} - A_{21}C_1^{-1}C_2)S_{22} + S_{22}(A_{22} - A_{21}C_1^{-1}C_2)^T, \tag{2.15}$$

where the matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

is partitioned in a similar form as Φ. It is noted that the minimum phase condition implies the stability of $A_{22} - A_{21}C_1^{-1}C_2$, because the matrix

$$\begin{bmatrix} sI - A_{11} & -A_{12} & I \\ -A_{21} & sI - A_{22} & 0 \\ -C_1 & -C_2 & 0 \end{bmatrix}\begin{bmatrix} I & -C_1^{-1}C_2 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} * & * & I \\ * & sI - A_{22} + A_{21}C_1^{-1}C_2 & 0 \\ -C_1 & 0 & 0 \end{bmatrix} \tag{2.16}$$

has full rank for all $\operatorname{Re} s \geqq 0$, if the system is minimum phase. Therefore, there does exist an $S_{22} > 0$, such that Φ_{22} in (2.15) is negative definite. In fact, (2.15) is simply a Lyapunov equation. Hence, by the minimum phase condition, we can take Φ_{22} as an arbitrary negative-definite matrix that will result in a positive-definite solution S_{22}. Clearly, if S_{22} is available, we can then obtain $S_{11}(S_{21} = S_{12}^T)$ by

$$S_{12} = -C_1^{-1}C_2S_{22}, \tag{2.17}$$

and S_{11} by the inequality

$$S_{11} > S_{12}S_{22}^{-1}S_{12}^T, \tag{2.18}$$

which ensures us that

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix}$$

is positive definite and symmetric. By (2.14), we may solve $\mathcal{K}$ by

$$\mathcal{K} = (C_1S_{11} + C_2S_{12}^T)^{-1}, \tag{2.19}$$

which is guaranteed to exist if S is positive definite. We now need to prove that with the matrix $\mathcal{K}$ and S obtained from (2.15), (2.17)-(2.19), there exists a $\rho > 0$, such that the matrix Φ defined in (2.12) is negative definite. Indeed, we have (derived from (2.12))

$$\Phi_{11} = A_{11}S_{11} + S_{11}A_{11}^T + S_{12}A_{12}^T + A_{12}S_{21} - 2\rho I_p, \tag{2.20}$$

which is negative definite if $\rho > 0$ is sufficiently large. The submatrix Φ_{22} is certainly negative definite, which was chosen to solve $S_{22} > 0$ in (2.15). It is noted that Φ_{12} is given by (also derived from (2.12))

$$\Phi_{12} = A_{11}S_{12} + A_{12}S_{22} + S_{11}A_{21}^T + S_{12}A_{22}^T, \tag{2.21}$$

which is independent of ρ. Therefore, if we choose $\rho > 0$ sufficiently large, (2.13) will be satisfied. Furthermore, the value of ρ can be computed by

$$\rho > \tfrac{1}{2}\sigma_{\max}(\Phi_{11} - \Phi_{12}\Phi_{22}^{-1}\Phi_{21} + 2pI), \tag{2.22}$$

with $\sigma_{\max}$ the maximum singular value. It is noted that the right-hand side above does not involve ρ by inspection of (2.15), (2.20), and (2.21). Hence, in light of Lemma 2.8, the quadratic cost functional $J(u)$ is minimized with output feedback gain $K = -\rho\mathcal{K}$ for $Q = -\Phi$ and $R = I_p/2\rho$.

Necessity. Suppose that the minimum phase condition is violated; then (2.15) can never have a solution $S_{22} > 0$, such that $\Phi_{22} \leqq 0$, by the Lyapunov Theorem with the assumption that the open-loop system (2.1) is absent of transmission zeros on an imaginary axis. The condition $\det(CB) \neq 0$ is also necessary, because if $\det(CB) = 0$, then the matrix $\mathcal{K}CB$ is singular, which contradicts the positive realness condition (see (2.9) with $F = KC$) by noting that the input matrix B has full column rank. □

The matrix CB is often called high-frequency gain. It should be emphasized that the above proof also provides a synthesis procedure (equations (2.15)-(2.22)) for the design of linear quadratic optimal regulator with linear, static output feedback compensator. Based on Theorem 2.10, we also obtain Corollary 2.23.

COROLLARY 2.23. *Let the open-loop system be defined as in* (1.1), *which is absent of transmission zeros on an imaginary axis. Suppose that the minimum phase or nonsingular high-frequency gain condition is violated; then there does not exist a dynamic output feedback compensator* $K(s)$ *such that the closed-loop system transfer function matrix* $K(s)C(sI - A + BK(s)C)^{-1}B$ *is positive real.*

Proof. Let the dynamic output feedback compensator be $K(s) = J + H(sI - F)^{-1}G$. Then the augmented open-loop system can be represented as

$$\begin{bmatrix} \dot{x} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} A & 0 \\ GC & F \end{bmatrix}\begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u, \qquad \hat{y} = [JC \quad H]\begin{bmatrix} x \\ z \end{bmatrix}. \tag{2.24}$$

Clearly, the unstable transmission zeros of the system $\{A, B, C\}$ are also the transmission zeros of the augmented system (2.24), and $\det([JC\,H]\begin{bmatrix} B \\ 0 \end{bmatrix}) = \det(JCB)$ because the open-loop system is a series connection of $\{A, B, C\}$ and $\{F, G, H, J\}$. Therefore, there does not exist a feedback law $u = -K\hat{y}$ such that the closed-loop system is positive real in light of Theorem 2.10. □

The above result tells us why the LTR (loop transfer recovery [2], [5]) technique does not work for nonminimum phase systems. In light of Theorem 2.10 and Corollary 2.23, the optimality property (infinite gain margin, sixty-degree phase margin for single-input systems or Kalman-Anderson inequality for multi-input systems) can never be recovered for nonminimum phase systems. It should be clarified that the optimality property for single-input systems is mainly determined by infinite gain margin, because if the system has infinite gain margin, the Nyquist plot of the loop

gain transfer function will eventually be outside of unit disc centered at $-1+j0$ in complex plane by increasing the feedback gain. Clearly, the above results also apply to nonsquare systems. This is because if the open-loop system is nonsquare, we may always find a static (or dynamic) compensator to square down the system while keeping the minimum phase property [9]. The details are omitted here.

It is noted that the presence of transmission zeros on an imaginary axis for system (1.1) does create a dilemma. It is possible that even if system (1.1) has transmission zeros on an imaginary axis, there exists an output feedback control $u=-Ky$ that minimizes $J(u)$ in (1.2) for some $R>0$ and $Q\geqq 0$ with $\{Q, A\}$ observable as illustrated below.

Consider system (1.1) with realization

$$A=\begin{bmatrix}3 & -2\\ 1 & 0\end{bmatrix},\quad B=\begin{bmatrix}1\\ 0\end{bmatrix},\quad C=[1\quad 0], \tag{2.25}$$

which has a transmission zero at the origin. If the output feedback law $u=-Ky$, with $K=10$, is used, the quadratic cost functional $J(u)$ in (1.2) is minimized for $R=\frac{1}{2}$, and $Q=\begin{bmatrix}20 & 0\\ 0 & 0\end{bmatrix}$. Indeed, the algebraic Riccati equation (2.2) yields a positive-definite solution $P=\begin{bmatrix}5 & 0\\ 0 & 10\end{bmatrix}$ such that $KC=R^{-1}B^TP$ is satisfied.

However, the fact illustrated in the above example is not true in general. Next consider a linear system with the following realization:

$$A=\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix},\quad B=\begin{bmatrix}1\\ 0\end{bmatrix},\quad C=[1\quad 0], \tag{2.26}$$

which also has a transmission zero at the origin. It can be easily verified that this system cannot even be stabilized by linear static output feedback control, and hence it does not admit an LQ optimal output feedback.

3. Application to the H_∞ optimal control. In this section, we consider a linear system described by the state equation

$$\dot{x}(t)=Ax(t)+Bu(t)+Dw(t),\qquad z(t)=Ex(t),\ y=Cx(t), \tag{3.1}$$

where $x(t)\in\mathbb{R}^n$ is the state, $w(t)\in\mathbb{R}^s$ is the disturbance input, $u(t)\in\mathbb{R}^m$ is the control input, $y(t)\in\mathbb{R}^p$ is the measured output, and $z\in\mathbb{R}^l$ is the controlled output. It is assumed that the number of control inputs m is the same as the number of measured outputs p.

The standard H_∞ optimization problem is concerned with constructing a dynamic feedback compensator $u=G(s)y$ to minimize the H_∞ norm of the transfer function from w to z (e.g., see [4]). A special case when the full state can be measured (i.e., $C=I$) is considered by Petersen [7]. The following notion is adopted here.

DEFINITION 3.2. Let the constant $\gamma>0$ be given. The system (3.1) is said to be stabilizable with disturbance attenuation γ if there exists a state feedback matrix $F\in\mathbb{R}^{m\times n}$ such that the following conditions are satisfied:

(1) The matrix $A-BF$ is a stability matrix. That is, all the eigenvalues of $A-BF$ lie in the open left half plane.

(2) The transfer function matrix

$$T(s)=E(sI-A+BF)^{-1}D \tag{3.3}$$

satisfies the bound $T^T(-j\omega)T(j\omega)\leqq\gamma^2 I$, for all $\omega\in\mathbb{R}$. That is, the H_∞ norm of $T(s)$ is less than or equal to γ: $\|T\|_\infty\leqq\gamma$.

When the states are not available for feedback, the dynamic output feedback compensator $G(s)$ may be used. Define

$$\gamma^* := \inf\{\|T_c\|_\infty : G(s) \text{ is proper}\}, \tag{3.4}$$

where $T_c(s) = E(sI - BG(s)C)^{-1}D$ is the transfer function for a closed-loop system. The task of H_∞ optimization is to synthesize a stabilizing dynamic output feedback compensator $G(s)$, which achieves γ^* in (3.4). A recent development in H_∞ control is that for any $\gamma > \gamma^*$, there exists a feedback compensator $G(s)$, which achieves disturbance attenuation γ, i.e., $\|T_c\|_\infty \leqq \gamma$. The remarkable feature of this result is [4] that the feedback compensator $G(s)$ can be obtained by solving two algebraic Riccati equations and $G(s)$ has the same McMillan degree as the open-loop system. Here, we will consider H_∞ optimization via linear, static output feedback $u = -Ky$, $K \in \mathbb{R}^{p\times p}$. Using the same technique as in § 2, we obtain the following result.

THEOREM 3.5. *Suppose that the realization $\{A, B, C\}$ is both stabilizable and detectable, and assume that the transfer function $C(sI - A)^{-1}B$ is strict minimum phase,* $\det(CB) \neq 0$, *and $D = B\Omega$ for some $\Omega \neq 0$. Then for any $\gamma > 0$, there exists a matrix $K \in \mathbb{R}^{p\times p}$ such that the closed-loop system is stabilized with attenuation γ, i.e., $\|T_c\|_\infty \leqq \gamma$.*

Proof. It has been shown in [7] that if the following algebraic Riccati equation:

$$A^TP + PA - \frac{1}{\varepsilon}PBB^TP + \frac{1}{\gamma}PDD^TP + \frac{1}{\gamma}E^TE + \varepsilon Q = 0, \tag{3.6}$$

has a positive-definite solution P for some positive-definite matrix Q and scalar $\varepsilon > 0$, then the state feedback control

$$u = -\frac{1}{2\varepsilon}B^TPx = -Fx, \tag{3.7}$$

stabilizes the closed-loop system with disturbance attenuation γ. If we consider the output feedback law $u = -(1/2\varepsilon)\mathcal{K}y$, then (3.7) reduces to

$$\frac{1}{2\varepsilon}B^TP = \frac{1}{2\varepsilon}\mathcal{K}C = KC. \tag{3.8}$$

Without loss of generality, we assume that $B^T = [I_p \;\; 0]$. Partition $C = [C_1 \;\; C_2]$, $C_1 \in \mathbb{R}^{p\times p}$ is nonsingular by condition $\det(CB) \neq 0$. Define

$$S := P^{-1} = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix} \quad \text{and partition } A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{3.9}$$

with S_{11}, $A_{11} \in \mathbb{R}^{p\times p}$. Now, multiplying (3.6) by $S = P^{-1}$ from both left and right, we obtain

$$SA^T + AS - B\left(\frac{I_p}{\varepsilon} - \frac{\Omega\Omega^T}{\gamma}\right)B^T + \frac{1}{\gamma}SE^TES + \varepsilon SQS = 0, \tag{3.10}$$

which is equivalent to the matrix

$$\Phi = SA^T + AS - B\left(\frac{I_p}{\varepsilon} - \frac{\Omega\Omega^T}{\gamma}\right)B^T + \frac{1}{\gamma}SE^TES < 0. \tag{3.11}$$

Therefore, with the conditions given, we need to find a positive-definite matrix S and a nonsingular matrix $\mathcal{K}$ and a scalar $\varepsilon > 0$ such that (3.8) and (3.11) are both satisfied. Clearly, (3.8) is equivalent to

(3.12) (i) $C_1S_{11} + C_2S_{21} = \mathcal{K}^{-1}$ and
(ii) $C_1S_{12} + C_2S_{22} = 0$.

Partition

$$\Phi=\begin{bmatrix}\Phi_{11} & \Phi_{12}\\ \Phi_{21} & \Phi_{22}\end{bmatrix} \quad \text{and} \quad E=[E_1 \quad E_2]$$

with $\Phi_{11}\in\mathbb{R}^{p\times p}$ and $E_1\in\mathbb{R}^{t\times p}$. Then we obtain

$$\begin{aligned}\Phi_{11}&=S_{11}A_{11}^T+A_{11}S_{11}+S_{12}A_{12}^T+A_{12}S_{21}-\left(\frac{I_p}{\varepsilon}-\frac{\Omega\Omega^T}{\gamma}\right)\\ &\quad+\frac{1}{\gamma}[S_{11}\quad S_{12}]E^TE[S_{11}\quad S_{12}]^T,\end{aligned} \tag{3.13}$$

$$\Phi_{22}=S_{22}(A_{22}-A_{21}C_1^{-1}C_2)^T+(A_{22}-A_{21}C_1^{-1}C_2)S_{22}+\frac{1}{\gamma}S_{22}\hat{E}^T\hat{E}S_{22} \tag{3.14}$$

with $\hat{E}=E_2-E_1C_1^{-1}C_2$, and

$$\Phi_{12}=S_{11}A_{21}^T+A_{11}S_{12}+S_{12}A_{22}^T+A_{12}S_{22}+\frac{1}{\gamma}[S_{11}\quad S_{12}]E^TE[S_{21}\quad S_{22}]^T. \tag{3.15}$$

Equations (3.13)-(3.15) are derived in a similar way as in § 2. By the strict minimum-phase condition, and by noting that $\Phi_{22}<0$ can be chosen arbitrarily, (3.14) does have a positive-definite solution S_{22}. Following the same procedure as in § 2, $S_{12}=S_{21}^T=-C_1^{-1}C_2S_{22}$ is obtained according to (3.12) and S_{11} is chosen to be such that $S_{11}>S_{12}S_{22}^{-1}S_{12}^T$, which ensures the positive definiteness of S. With S_{ij} so determined, we can make Φ_{11} in (3.13) negative definite provided that $\varepsilon>0$ is sufficiently small. Finally, by noting that Φ_{12} and Φ_{22} in (3.14), (3.15) do not involve ε, we can thus choose an $\varepsilon>0$ such that $\Phi_{11}-\Phi_{22}\Phi_{22}^{-1}\Phi_{21}<0$, which guarantees that the matrix Φ is negative definite. Therefore, by taking $P=S^{-1}$, and $K=\mathcal{K}/2\varepsilon$, where $\mathcal{K}$ is solved from (3.12), the transfer function $T_c(s)=E(sI-A+BKC)^{-1}D$ for a closed-loop system does satisfy $\|T_c\|_\infty\leqq\gamma$. □

It should be clarified that the result presented in Theorem 3.5 is a direct consequence of Theorem 2.10. However, the conditions imposed on system (3.1) may not be necessary for the existence of H_∞ optimal control with linear, static output feedback because of the additional constraint $D=B\Omega$. This constraint is not needed if a dynamic output feedback compensator is used. It has been shown by Petersen and Hollot [8] that as long as the transfer function $C(sI-A)^{-1}D$ is minimum phase, matrices C, D are of full rank, $\{A, D\}$ stabilizable, $\{C, A\}$ detectable, and there exists an nth order dynamic output feedback compensator $G(s)$ such that system (3.1) is stabilized with the same disturbance attenuation γ as achievable by full state feedback. It is not known at present if the result in [8] is also true for linear, static output feedback compensators.

4. An example. Let the linear system be given as in (1.1) with realization

$$A=\begin{bmatrix}0&0&0&0\\0&0&0&0\\2&0&1&2\\0&2&2&1\end{bmatrix},\quad B=\begin{bmatrix}1&0\\0&1\\0&0\\0&0\end{bmatrix},\quad C^T=\begin{bmatrix}4&-2\\-3&2\\3&-1\\0&2\end{bmatrix}, \tag{4.1}$$

where $p=m=2$. Clearly, $\det(CB)\neq 0$. Considering that $B^T=[Ip\ \ 0]^T$, we find that the matrix (with the same partition as in § 2)

$$A_{22}-A_{21}C_1^{-1}C_2=\begin{bmatrix}1&2\\2&1\end{bmatrix}-\begin{bmatrix}2&0\\0&2\end{bmatrix}\begin{bmatrix}4&-3\\-2&2\end{bmatrix}^{-1}\begin{bmatrix}3&0\\-1&2\end{bmatrix}=\begin{bmatrix}-2&-4\\0&-7\end{bmatrix} \tag{4.2}$$

is stable. Hence, the system $\{A, B, C\}$ is minimum phase. We may directly solve matrices S and $\mathcal{K}$ as in (2.11) following the equations (2.14)-(2.19). $S_{22}=\begin{bmatrix}9.875 & -1\\ -1 & 2.250\end{bmatrix}$ is obtained by solving the Lyapunov equation

$$\Phi_{22}=-\begin{bmatrix}31.5 & 0\\ 0 & 31.5\end{bmatrix}=(A_{22}-A_{21}C_1^{-1}C_2)S_{22}+S_{22}(A_{22}-A_{21}C_1^{-1}C_2)^T. \tag{4.3}$$

$S_{12}=\begin{bmatrix}-11.81 & -5.250\\ -5.875 & -8.000\end{bmatrix}$ is obtained by (2.17).

$$S_{11}=10I_2+S_{12}S_{22}^{-1}S_{12}^T=\begin{bmatrix}43.47 & 32.81\\ 32.81 & 47.88\end{bmatrix}>0$$

and $\mathcal{K}=\begin{bmatrix}0.100 & 0.150\\ 0.100 & 0.200\end{bmatrix}$ are also obtained from (2.18), (2.19). From (2.20) and (2.21), we have that

$$\Phi_{11}=-2\rho I_2 \quad \text{and} \quad \Phi_{12}=\begin{bmatrix}64.63 & 36.75\\ 43.75 & 76.00\end{bmatrix} \tag{4.4}$$

are obtained from (2.21). Next, we use the following inequality to determine ρ in (2.22):

$$\rho>\tfrac{1}{2}\sigma_{\max}(\Phi_{11}-\Phi_{12}\Phi_{22}^{-1}\Phi_{21}+2\rho I). \tag{4.5}$$

Inequality (4.5) is satisfied if $\rho>198.6$. Hence, by taking $\rho=200$, the final required output feedback gain is $K=\rho\mathcal{K}=\begin{bmatrix}20 & 30\\ 20 & 40\end{bmatrix}$. With $P=S^{-1}$, $\mathcal{K}$ and ρ obtained above, we find that

$$(A-\rho B\mathcal{K}C)^TP+P(A-\rho B\mathcal{K}C)<0 \quad \text{and} \quad \mathcal{K}C=B^TP, \tag{4.6}$$

are satisfied, which verifies the optimality of the output feedback $u=-Ky$ with respect to the cost functional $J(u)$ in (1.2) for $Q=-\Phi$ and $R=I_2/400$.

5. Conclusion. We have studied an optimal linear quadratic regulator problem with static output feedback. A necessary and sufficient condition is established for the existence of optimal output feedback control with respect to the cost functional (1.2) under the assumption that the system (1.1) does not have transmission zeros on an imaginary axis. The presence of transmision zeros on an imaginary axis does bring some difficulties in studying the existence of optimal output feedback. However, it is conjectured that as long as system (1.1) is stabilizable and detectable, and is void of transmission zeros on the open right half plane (minimum phase condition), there exists a dynamic output feedback compensator such that the optimality property for the loop transfer function can be recovered. We use the following example to conclude our paper.

Example 5.1. Consider plant $P(s)=C(sI-A)^{-1}B$ with realization $\{A, B, C\}$ as in (2.26). It has been shown earlier that such a system cannot even be stabilized by constant output feedback. If we use dynamic output feedback compensator $K(s)=\rho(s+1)/(s-2)$, then the loop transfer function $L(s)=K(s)P(s)=H(sI-F)^{-1}G$ has a minimal realization

$$F=\begin{bmatrix}3 & -2\\ 1 & 0\end{bmatrix}, \quad G=\begin{bmatrix}1\\ 0\end{bmatrix}, \quad H=\rho[1 \quad 0], \tag{5.2}$$

which is exactly the same as the realization (2.25) except for parameter ρ. Hence, if we choose $\rho = 10$, the optimality property (infinite gain margin and sixty-degree phase margin) can be achieved as shown in § 2.

Acknowledgments. The author thanks the anonymous reviewers for their helpful comments.

REFERENCES

[1] B. D. O. ANDERSON AND S. VONGPANITLERD, *Network Analysis and Synthesis—A Modern System Theory Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[2] J. C. DOYLE AND G. STEIN, *Multivariable feedback design: concepts for a classical/modern synthesis*, IEEE Trans. Automat. Control, 26 (1981), pp. 4-16.

[3] B. A. FRANCIS, *The optimal linear-quadratic time-invariant regulator with cheap control*, IEEE Trans. Automat. Control, 24 (1979), pp. 616-621.

[4] K. GLOVER AND J. C. DOYLE, *State-space formulae for all stabilizing controllers that satisfy an H_∞-norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167-172.

[5] R. E. KALMAN, *When is a linear control system optimal?*, Trans. ASME Ser. D. J. Basic Engrg., 86 (1964), pp. 51-60.

[6] N. A. LEHTOMATI, N. R. SANDELL, AND M. ATHANS, *Robustness results in linear-quadratic Gaussian based multivariable control designs*, IEEE Trans. Automat. Control, 26 (1981), pp. 75-93.

[7] I. R. PETERSEN, *Disturbance attenuation and H^∞ optimization: design method based on the algebraic Riccati equation*, IEEE Trans. Automat. Control, 32 (1987), pp. 427-429.

[8] I. R. PETERSEN AND C. V. HOLLOT, *High gain observers applied to problems in the stabilization of uncertain systems, disturbance attenuation and H^∞ optimization*, in Proc. American Control Conference, Atlanta, GA, 1988, pp. 2490-2496; Internat. J. Adaptive Control and Signal Process., to appear.

[9] A. SABERI AND P. SANNUTI, *Squaring down by static and dynamic compensators*, IEEE Trans. Automat. Control, 33 (1988), pp. 358-365.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 720-732, May 1990

014

OPTIMAL CONTROL WITH SMALL GENERALIZED GRADIENTS*

JAY S. TREIMAN†

Abstract. One of the main uses of generalized gradients is obtaining tight optimality conditions for optimal control problems. In this paper it is shown that the B-gradients satisfy the formula stating that the generalized gradients of an integral functional are contained in the "integral" of the generalized gradients. This formula is then applied to derive Euler-Lagrange equations and optimality conditions for a differential inclusion problem. All of these conditions can be stated in simple forms.

Key words. optimal control, Euler-Lagrange equations, differential inclusions, B-gradients, generalized gradients

AMS(MOS) subject classifications. 49B99, 58C20

1. Introduction. The most difficult and interesting area of application for generalized gradients is optimal control. The reason for the difficulties is that formulas for the generalized gradients of integral functionals are not easily derived. The other side of this is the possibility of elegant optimality conditions that subsume many classical results.

It has been clearly demonstrated by Ward and Merkovsky [11], [12] that any notion of directional derivative meeting fairly minimal conditions yields very general optimality conditions. Unfortunately, this type of general optimality condition usually does not apply directly to optimal control problems.

A very good example of this comes from the following basic proposition of Clarke. Here $d(C, x)$ is the distance from X to the set C.

PROPOSITION 1.1 [3, Prop. 2.4.3]. *Let X be a Banach space, C a closed subset of X, and f a Lipschitz function from X to $\mathfrak{R}$. If $\mathbf{x}$ is the minimum of f relative to C, then, for some $k > 0$,*

$$f(x) + kd(C, x)$$

has its global minimum at $\mathbf{x}$.

From this lemma, if $\partial_A g(x)$ denotes the set of generalized gradients related to a given tangent cone, and the A-gradients have a decent calculus, we can derive the optimality condition:

$$0 \in \partial_A f(\mathbf{x}) + N_A(C, \mathbf{x}).$$

Here $N_A(C, \mathbf{x})$ is the polar of the A-tangent cone to C at $\mathbf{x}$.

This is a simple and useful optimality condition. It is, however, useless in optimal control without results relating the generalized gradients of an integral functional to the integral of generalized gradients. The major success in this area is the generalized gradient definition of Clarke. Applying the above optimality condition in a variety of ways, Clarke derives elegant optimality conditions for optimal control problems (see Chapters 3, 4, and 5 of [3]).

The main objective of this paper is to show that we can obtain "nice" optimality conditions for optimal control problems using generalized gradient sets that are smaller than those of Clarke. This can eliminate spurious solutions from consideration. Here the B-gradients are the set of generalized gradients used (see [4], [6], [7], [9], and [10]).

* Received by the editors March 13, 1989; accepted for publication (in revised form) August 7, 1989.
† Department of Mathematics and Statistics, Western Michigan University, Kalamazoo, Michigan 49008.

The remainder of this paper is divided into four sections: the properties of the B-gradients are discussed in § 2; a formula for the B-gradient of integral functionals is derived in § 3; an application of the results in § 3 to the Euler-Lagrange equations is given in § 4; and § 5 demonstrates the use of B-gradients for a differential inclusion problem.

2. Preliminaries. When defining the B-gradients, the main objective was to find a convex set of generalized gradients that is smaller than Clarke's gradients while maintaining the utility of Clarke's gradients [1]-[3]. The B-gradients were defined in [9] and [4] through a tangent cone that is a modification of the cone of [7]. It should be noted that these two tangent cones are identical in finite-dimensional spaces and share some properties. Even so, there are differences. The major difference is in the variety of characterizations of the B-tangent cone that are not available for the Penot tangent cone.

The B-gradients are defined in a manner similar to the Clarke gradients and have a similar number of different characterizations [9], [4]. Here, both of these are defined in terms of directional derivatives.

Throughout the rest of this section E will be a Banach space.

DEFINITION 2.1. Let f be a lower semicontinuous (l.s.c.) function from a Banach space E to $\mathfrak{R}$. The Clarke derivative [8], [3] of f at x in the direction h is

$$f^{\uparrow}(x; h) = \sup_{\varepsilon>0} \limsup_{\substack{t\to 0^+ \\ x'\to x}} \inf_{h'\in B(h,\varepsilon)} \frac{f(x'+th')-f(x')}{t}.$$

The B-derivative [9] of f at x in the direction h is

$$f^{B}(x; h) = \sup\{\alpha, \beta\},$$

where

$$\alpha = \sup_{\substack{\lambda>0 \\ \varepsilon>0}} \limsup_{x'\to x,\, x'\neq x} \inf_{h'\in B(h,\varepsilon)} \frac{f(x'+\lambda\|x'-x\|_f h')-f(x')}{\lambda\|x'-x\|_f}$$

and

$$\beta = \liminf_{t\to 0^+,\, y'\to y} \frac{f(x+ty')-f(x)}{t}.$$

Here $\|x'-x\|_f = \|x'-x\| + |f(x)-f(x')|$.

In the above, β is the lower Dini derivative, $f^{K}(x; h)$. When dealing with Lipschitz functions the above definitions can be simplified.

THEOREM 2.2 [8], [9]. *If $f: E\to\mathfrak{R}$ is Lipschitz, and x and h are in E, then*

$$f^{\uparrow}(x; h) = f^{0}(x; h) = \limsup_{x'\to x} \frac{f(x'+th)-f(x')}{t}$$

and

$$f^{B}(x; h) = \sup\{\alpha, \beta\}$$

where

$$\alpha = \sup_{\lambda>0} \limsup_{x'\to x} \frac{f(x'+\lambda\|x'-x\|h)-f(x')}{\lambda\|x'-x\|}$$

and

$$\beta = \liminf_{t\to 0^+,\, y'\to y} \frac{f(x+ty')-f(x)}{t}.$$

In the above theorem the lower Dini derivative can be replaced with the upper Dini derivative,

$$f^k(x; h) = \limsup_{t \to 0^+,\, y' \to y} \frac{f(x+ty')-f(x)}{t}.$$

This definition shows the relationship between the two derivatives. The B-derivative is less than the Clarke derivative because of the restriction on the "t" values.

In this work we will need an alternate definition of $f^B(x; h)$. This variant is used to prove several results. Under the hypothesis that f is Lipschitz,

$$f^B(x; h) = \sup_{\lambda > 0} \limsup_{\substack{t \to 0^+ \\ x' \to x}} \inf_{t \in (0, \lambda \|x'-x\|]} \frac{f(x'+th)-f(x')}{t}.$$

Using these definitions we define the corresponding generalized gradients.

DEFINITION 2.3. Let f be as above. The Clarke generalized gradients (CGG), $\partial f(x)$, to f at x are the set of $x^* \in E^*$ such that

$$f^{\uparrow}(x; h) \geqq (x^*, h) \quad \text{for all } h \in E.$$

The B-gradients, $\partial_B f(x)$, to f at x are the set of $x^* \in E^*$ such that

$$f^B(x; h) \geqq (x^*, h) \quad \text{for all } h \in E.$$

The above observation on the relationship between the directional derivatives yields the following lemma relating these generalized gradients.

LEMMA 2.4 [9]. *Let $f: E \to \Re$ be lower semicontinuous and x a point where f is finite. Then*

$$\partial_B f(x) \subset \partial f(x).$$

In extending the results of Clarke, the main gain is in the tightness of the results. One simple way to show this is through the corollaries in later sections based on the following result.

THEOREM 2.5. *If f is strictly differentiable* [3, Prop. 2.2.1] *at x then*

$$\partial f(x) = \{\nabla f(x)\}.$$

If f is Frechet differentiable at x, then

$$\partial_B f(x) = \{\nabla f(x)\}.$$

Proof. The first half of the proof is Proposition 2.2.1 of [3]. The second statement is an exercise in comparing the definitions of $\partial_B f(x)$ and the Frechet derivative. □

A simple function demonstrating this difference on $\Re$ is

$$g(x) = \begin{cases} x^2 \sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Here g is differentiable at zero. Thus $\partial_B g(0) = \{0\}$, but $\partial f(0) = [-1, 1]$.

Using these definitions we can prove a variety of calculus results for these generalized gradients. All of the results in this section are stated for the B-gradients, but the same results hold for CGG. The first result is a simple sum formula.

THEOREM 2.6 [9]. *Let $f: X \to \Re$ be lower semicontinuous with $f(\bar{x})$ finite and let $g: X \to \Re$ be Lipschitz. Then*

$$\partial_B(f+g)(\bar{x}) \subset \partial_B f(\bar{x}) + \partial_B g(\bar{x}).$$

This is equivalent to

$$(f+g)^B(\bar{x}; h) \leqq f^B(\bar{x}; h) + g^B(\bar{x}; h).$$

The next result gives us another way to calculate the B-gradients of a sum of functions in a special case. This is very important in proving results in optimal control.

THEOREM 2.7. *For two Banach spaces X and Y let $f: X \to \Re$ and $g: Y \to \Re$ be Lipschitz functions. If $F: X \times Y \to \bar{\Re}$ is given by $F(x, z) = f(x) + g(z)$, then*

$$\partial_B F(\mathbf{x}, \mathbf{z}) \subset \partial_B f(\mathbf{x}) \times \partial_B g(\mathbf{z}).$$

Proof. Assume that f and g have Lipschitz constant L, $(x, z) = (0, 0)$, and fix $\lambda > 0$.

We will show that if $\omega > f^B(x; h)$ and $\eta > g^B(z; k)$, then $F^B((x, z); (h, k)) \leqq \omega + \eta$. We only need to consider the α part of the definition of $F^B((x, z); (h, k))$ (see Definition 2.1 above) since this result is known for lower Dini derivatives. Thus we assume that all (v, w) under consideration are not $(0, 0)$.

Our goal is accomplished by demonstrating that there are neighborhoods N of x and M of z such that if $(v, w) \in N \times M$, then

$$\frac{F((v, w) + \lambda \|(v, w)\|_F \cdot (h, k)) - f(v, w)}{\lambda \|(v, w)\|_F} \leqq \omega + \eta. \tag{2.1}$$

If h or k is zero, its component drops from (2.1) and can be ignored in the rest of the proof.

The case of $(v, w) = (0, 0)$ follows from the calculus of the directional derivative corresponding to the contingent cone. Thus we assume that $(v, w) \neq (0, 0)$.

Fix $\tau > 0$, $\omega' < \omega$ and $\eta' < \eta$ such that

$$\left(1 + \frac{\tau}{\lambda}\right)(\omega' + \eta') + L\tau(\|h\| + \|k\|) < \omega + \eta.$$

There are neighborhoods $N = B(0, r)$ of zero in X and $M = B(0, r)$ of zero in Y such that for all $v' \in 2N$ and $w' \in 2M$,

$$\frac{f(v' + \tau\|v'\|h) - f(v')}{\tau\|v'\|} < \omega'$$

and

$$\frac{f(w' + \tau\|w'\|h) - f(w')}{\tau\|w'\|} < \eta'.$$

For both v and w there are two cases to be considered: when the component is nonzero and when the component is zero. Assume $v \neq 0$. Then $\|v\| \leqq \|(v, w)\|_F$ and there is a sequence $v = v_0, v_1, \cdots, v_m$ such that

$$v_{i+1} = v_i + \tau\|v_i\|h, \qquad i = 0, 1, 2, \cdots, m-1, \tag{2.2}$$

$$\|v_m - v\| - \lambda\|(v, w)\|_F\|h\| < \tau\|(v, w)\|_F\|h\|, \tag{2.3}$$

and

$$\frac{f(v_{i+1}) - f(v_i)}{\|v_{i+1} - v_i\|}\|h\| = \frac{f(v_i + \tau\|v_i\|h) - f(v_i)}{\|v_{i+1} - v_i\|} < \omega', \qquad i = 0, 1, 2, \cdots, m-1. \tag{2.4}$$

This implies that

$$\begin{aligned}\omega' &> \frac{f(v_m) - f(v)}{\|v_m - v\|}\|h\| \\ &> \frac{f(v + \lambda\|(v, w)\|_F h) - f(v) + L\tau\|(v, w)\|_F\|h\|}{(\lambda\|(v, w)\|_F + \tau\|(v, w)\|)\|h\|}\|h\|,\end{aligned}$$

and hence

$$\frac{f(v+\lambda\|(v,w)\|_F h)-f(v)}{\lambda\|(v,w)\|_F}<\left(1+\frac{\tau}{\lambda}\right)\omega'+L\tau\|h\|. \tag{2.5}$$

If $v=0$, because for some γ arbitrarily close to zero

$$\frac{f(\gamma h)-f(0)}{\gamma}<\omega',$$

there is a sequence satisfying (2.2) for $i=1,2,\cdots,m-1$, (2.3), and (2.4) for $i=0,1,\cdots,m-1$ where $v_1=\gamma h$. This implies that (2.5) also holds if $v=0$.

The same argument applied to the Y component yields

$$\frac{g(w+\lambda\|(v,w)\|_F k)-g(w)}{\lambda\|(v,w)\|_F}<\left(1+\frac{\tau}{\lambda}\right)\eta'+L\tau\|k\|$$

and finally

$$\frac{F((v,w)+\lambda\|(v,w)\|_F\cdot(h,k))-f(v,w)}{\lambda\|(v,w)\|_F}<(1+\tau)(\omega'+\eta')+L\tau(\|h\|+\|k\|)$$
$$<\omega+\eta. \qquad \square$$

To translate between set constraints and generalized gradients, a result relating the B-gradients of the distance function from a set and the B-normal cone is required.

DEFINITION 2.8. Let C be a closed subset of a Banach space E. The B-tangent cone, $T_B(C,x)$, to C at $\mathbf{x}$ is the set of all y in E such that

$$\forall\varepsilon>0,\quad \lambda>0\quad \exists N\in\mathcal{N}(\mathbf{x})\quad \forall x'\in C\backslash\mathbf{x},$$
$$C\cap\{x'+\lambda\|x'-\mathbf{x}\|B(y,\varepsilon)\}\neq\emptyset$$

and for all $\lambda>0$ and $\varepsilon>0$,

$$C\cap\{x'+(0,\lambda]B(y,\varepsilon)\}\neq\emptyset.$$

The B-normal cone, $N_B(C,\mathbf{x})$ to C at $\mathbf{x}$ is the set of $x^*\in E^*$ such that

$$(x^*,y)\leqq 0$$

for all $y\in T_B(C,x)$.

The next result is used in restating optimality conditions.

THEOREM 2.9. *If C is a closed subset of X and $x\in C$, then*

$$\bigcup_{\lambda\geqq 0}\lambda\partial_B d(C,x)=N_B(C,x).$$

3. B-gradients and integral functionals. The result in this section gives the relationship between the B-gradients of an integral functional and the integral of the B-gradients of the function being integrated.

We will assume that $(T,\mathcal{T},\mu)$ is a positive real measure space with $\mu(T)<\infty$ and that Y is a separable Banach space. Let X be a closed subspace of $L^p(T,\Re^n)$ for some $p\in[0,\infty]$. Let $f_t:\Re^n-\Re$ be a given family of functions such that for each $y\in\Re^n$ the function $t:T\to f_t(y)$ is measurable, and that

$$F(x)=\int_T f_t(x(t))\mu(dt)$$

is defined and finite at x.

The following hypothesis is also assumed.

HYPOTHESIS 3.1. There is a function k in $L^q(T, \Re)$ such that, for all $t \in T$,

$$|f_t(y_1) - f_t(y_2)| \leqq k(t)\|y_1 - y_2\|$$

for all $y_1, y_2 \in \Re^n$. Here $1/p + 1/q = 1$ ($q = \infty$ if $p = 1$ and $q = 1$ if $p = \infty$).

The main result on the B-gradients of integral functionals is as follows.

THEOREM 3.2. *Under the above hypotheses, F is Lipschitz on bounded subsets of X, and*

$$\partial_B F(x) \subset \int_T \partial_B f_t(x(t))\mu(dt).$$

This is interpreted as follows. If $\psi \in \partial_B F(x)$, as a function of x, then $\psi(t)$, as a function of t, is in $L^1(T, \Re^n)$ and

$$\psi(t) \in \partial_B f_t(x) \quad \mu\text{-a.e.}$$

and, if $v \in L^p(T, \Re^n)$,

$$\langle \psi, v \rangle = \int_T \langle \psi(t), v(t) \rangle \mu(dt).$$

Proof. This result was proven for the case of $p = \infty$ in [9].

The proof that F is Lipschitz is contained in [3, § 2.7].

We assume that $\bar{x}$ is zero. For each $h \in X$ we will show that

$$F^B(0; h) \leqq \int_T f_t^B(0, h(t))\mu(dt).$$

Let

$$\begin{aligned} M &= F^B(0; h) \\ &= \max\left\{ \sup_{\lambda > 0} \limsup_{x \to 0} \int_I \frac{f_t(x(t) + \lambda\|x\|h(t)) - f_t(x(t))}{\lambda\|x\|} \mu(dt), F^k(x; h) \right\}. \end{aligned}$$

From the definition of $F^B(0, h)$, either for every $\varepsilon > 0$ there are $\lambda > 0$ and a sequence $x^k \to 0$ such that

$$M - \varepsilon < \limsup_{k \to \infty} \int_T \frac{f_t(x^k(t) + \lambda\|x^k\|h(t)) - f_t(x^k(t))}{\lambda\|x^k\|} \mu(dt), \tag{3.1}$$

or $F^k(0; h) = M$.

Using this result for the upper Dini derivative, the second case follows easily:

$$\begin{aligned} M &= F^k(0; h) \\ &\leqq \int_T f_t^k(0; h(t))\mu(dt) \\ &\leqq \int_T f_t^B(0; h(t))\mu(dt). \end{aligned}$$

In the first case we modify the difference quotients in the integral (1) as follows. There is a constant α such that for any $x \in L^p[0, 1]$, if $J = \{t: x(t) > \alpha\|x\|\}$, then

$2\int_J L(t)\mu(dt) < \varepsilon$. For each k replace the difference quotient by

$$g_k(t) = \begin{cases} \dfrac{f_t(x^k(t) + \lambda \|x^k\| h(t)) - f_t(x^k(t))}{\lambda \|x^k\|} & \text{if } x^k(t) \leqq \alpha \|x^k\|, \\ -L(t) & \text{if } x^k(t) > \alpha \|x^k\|. \end{cases}$$

It is clear that for each k this is a measurable function bounded in norm by $L(t)$ for each t. Then, applying Fatou's Lemma,

$$M - 2\varepsilon \leqq \limsup_{k\to\infty} \int_T g_k(t)\mu(dt)$$

$$\leqq \int_T \limsup_{k\to\infty} g_k(t)\mu(dt).$$

We only need to show that $\delta_t = \limsup_{k\to\infty} g_k(t) \leqq f_t^B(0; h(t))$ for all t. Fix t and let $\{k_i\}$ be a subsequence of the k's such that $\lim_{k_i\to\infty} g_k(t) = \delta_t$ and $|x^{k_i}(t)|/\|x^{k_i}\|$ converges to some ϕ. If ϕ is zero, then a simple argument shows that

$$\delta_t \leqq f_t^k(0; h(t)) \leqq f_t^B(0; h(t)).$$

If ϕ is not zero then, for some λ_0,

$$\delta_t = \lim_{k_i\to\infty} \frac{f_t(x^{k_i}(t) + \lambda_0 |x^{k_i}(t)| h(t)) - f_t(x^k(t))}{\lambda_0 |x^{k_i}|}$$

$$\leqq f_t^B(0; h(t)).$$

Thus

$$M \leqq 2\varepsilon + \int_T f_t^B(0; h(t))\mu(dt).$$

We therefore conclude that

$$F^B(0; h) \leqq \int_T f_t^B(0; h(t))\mu(dt).$$

Since $\partial F^B(0; \cdot) = \partial_B F(0)$ and $\partial f_t^B(0; h(t)) = \partial_B f_t(0)$, Theorem 1 [5, p. 13] and Theorem 1 [5, p. 20] yield the desired result. □

At this point it is appropriate to note that the above result also holds for the generalized gradients defined through the tangent cone of Michel and Penot [6], [7]. This follows from the fact that the B-tangent cone and the Penot tangent cone are the same in $\mathfrak{R}^n$ but the B-tangent cone is always contained in the Penot cone [4], [9].

COROLLARY 3.3. *Let $\partial_P f(x)$ denote the set of subgradients associated with the tangent cone of Michel and Penot. Then, if f satisfies the hypotheses of this section,*

$$\partial_P F(x) \subset \int_T \partial_P f_t(x(t))\mu(dt) = \int_T \partial_B f_t(x(t))\mu(dt).$$

This result shows that in optimal control sharper results cannot be obtained using the tangent cone of Michel and Penot instead of the B-tangent cone.

4. The Euler–Lagrange equations. We consider the following optimal control problem where the constraint does not depend on the state. Here $x: [a, b] \to \mathfrak{R}^n$ with a and b finite and $U(t)$ is a measurable multifunction from $[a, b]$ to $\mathfrak{R}^n$ with closed values. It is also assumed that $f(t, x)$ is Lipschitz with the same Lipschitz constant L for each t and $f(t, x)$ is measurable for each absolutely continuous $x(t)$.

The objective is to minimize

$$(P_1) \qquad \int_a^b f(t, x)\, dt$$

subject to

$$\dot{x} \in U(t), \quad x(a) = x_0, \quad x(b) = x_1.$$

Our basic result is the following.

THEOREM 4.1. *If* $\mathbf{x}$ *is a solution of* (P_1) *then there is an absolutely continuous curve* $p : [a, b] \to \Re^n$ *such that for all* $t \in [a, b]$,

$$p(t) \in N_B(U(t), \dot{\mathbf{x}}(t))$$

and

$$\dot{p}(t) \in \partial_B f(t, \mathbf{x}(t)).$$

Proof. We assume that $\mathbf{x}(t) = 0$. By Proposition 1.1 if S is the set of feasible arcs, a nonempty closed set, then the functional

$$\int_a^b f(t, x)\, dt + L' d(x, S)$$

has its unconstrained minimum at $\mathbf{x}$ over the set of absolutely continuous arcs. Here $L' \geqq L(b-a)$. Over these arcs we have the inequality

$$d(x, S) \leqq \int_a^b d(\dot{x}, U(t))\, dt + d(x(a), x_0) + d(x(b), x_1).$$

Thus $\mathbf{x}$ is the unconstrained minimum of

$$(4.1) \qquad F(x) = \int_a^b [f(t, x) + Ld(\dot{x}, U(t))]\, dt + L(d(x(a), x_0) + d(x(b), x_1))$$

over the set of all absolutely continuous arcs.

Taking the subgradient of F as a function on the set

$$X = \Big\{ (v, w, \alpha, \beta) \in L^\infty[a, b] \times L^\infty[a, b] \times \Re^n \times \Re^n : \exists c \in \Re^n \text{ with } v(t) = c + \int_a^t w(s)\, ds \Big\}$$

and applying Theorem 2.7, we get

$$\begin{aligned} (0, 0, 0, 0) &\in \partial_B F \\ &\subset \partial_B \int_a^b [f(t, \mathbf{x}(t)) + L' d(\dot{\mathbf{x}}(t), U(t))]\, dt \\ &\quad \times L' \partial_B d(\mathbf{x}(a), x_0) \times L' \partial_B d(\mathbf{x}(b), x_1) \\ &\subset \int_a^b \partial_B [f(t, \mathbf{x}(t)) + L' d(\dot{\mathbf{x}}(t), U(t))]\, dt \\ &\quad \times L' \partial_B d(\mathbf{x}(a), x_0) \times L' \partial_B d(\mathbf{x}(b), x_1). \end{aligned}$$

The conditions that $0 \in L'\partial_B d(\mathbf{x}(a), x_0)$ and $0 \in L'\partial_B d(\mathbf{x}(b), x_1)$ are simply the conditions that $\mathbf{x}(a) = x_0$ and $\mathbf{x}(b) = x_1$. The equation

$$(0, 0) \in \int_a^b [\partial_B f(t, \mathbf{x}(t)) + \partial_B L'd(\dot{\mathbf{x}}(t), U(t))]\, dt$$

implies that there are $(p, q) \in L^\infty[a, b] \times L^\infty[a, b]$ such that $q = \dot{p}$,

$$p(t) \in \partial_B L'd(\dot{\mathbf{x}}, U(t)) \subset N_B(U(t), \dot{\mathbf{x}}(t)), \qquad \mu\text{-a.e.},$$

and

$$q(t) \in \partial_B f(t, \mathbf{x}(t)), \qquad \mu\text{-a.e.} \qquad \square$$

It is interesting to compare this with classical results. In the classical results it is assumed that f is "smooth" in the x variable. Here we only need assume that f is "differentiable" to get the equality of the classical results.

COROLLARY 4.2. *Let* $\mathbf{x}$ *be a solution to* (P_1). *If* $f(t, x)$ *is differentiable, in the Fréchet sense, at* $\mathbf{x}(t)$ *for each* t, *and* $U(t)$ *is given by*

$$U(t) = \{x\colon g_t(x) \leqq 0\},$$

where g_t *is continuously differentiable at* $\mathbf{x}(t)$ *for each* t *with nonzero derivative, then there exists an absolutely continuous* p *such that for some* $l_0(t)$

$$p(t) = l_0(t)\nabla g_t(\mathbf{x}(t)) \quad \textit{and} \quad \dot{p}(t) = \nabla_x f(t, \mathbf{x}(t)).$$

A simple example shows that the number of functions satisfying these optimality conditions can be reduced in comparison with those given using the Clarke gradients.

Example 4.3. In the above problem let $U\colon [0, 1] \to \mathfrak{R}^2$ be the cardioid given by

$$U(t) = \begin{cases} x_1^2 + (x_2 - \frac{1}{2})^2 \leqq \frac{1}{4} & \text{if } x_1, x_2 \geqq 0, \\ x_1^2 + (x_2 + \frac{1}{2})^2 \leqq \frac{1}{4} & \text{if } x_1 \geqq 0, x_2 \leqq 0, \\ x_1^2 + x_2^2 \leqq 1 & \text{if } x_1 \leqq 0, \end{cases}$$

for each t, and let

$$f(t, (x_1, x_2)) = \begin{cases} 2x_2 & \text{if } t \in [0, \frac{1}{2}], \\ -x_2 & \text{if } t \in (\frac{1}{2}, 1]. \end{cases}$$

The endpoint constraints are $x(0) = x(1) = (0, 0)$. The subgradients of f are

$$\partial f(t, (x_1, x_2)) = \partial_B f(t, (x_1, x_2)) = \begin{cases} (0, -2) & \text{if } t \in (0, \frac{1}{2}), \\ (0, 1) & \text{if } t \in (\frac{1}{2}, 1). \end{cases}$$

The only pair of functions that satisfy the conditions of Theorem 4.1 are

$$\mathbf{p}(t) = \begin{cases} (0, -2t) & \text{if } t \in [0, \frac{1}{2}], \\ (0, t - \frac{3}{2}) & \text{if } t \in (\frac{1}{2}, 1], \end{cases}$$

and

$$\mathbf{x}(t) = \begin{cases} (0, -t) & \text{if } t \in [0, \frac{1}{2}], \\ (0, t - 1) & \text{if } t \in (\frac{1}{2}, 1]. \end{cases}$$

Thus $\mathbf{x}$ is the optimal arc.

If the conditions are written in terms of the Clarke gradients and normal cone, the function $\mathbf{x}(t) = (0, 0)$ also satisfies the optimality conditions. This is due to the fact that $N_{U(t)}(0, 0) = 0 \times \mathfrak{R}$. The same p function works for this $\mathbf{x}$ under these conditions.

It should be noted that this example is typical of situations where the B-gradients are an improvement over the Clarke gradients.

5. Differential inclusions. The basic differential inclusion problem studied here is minimizing a function $f(b)$ subject to

$$(P_2) \qquad \dot{x}(t) \in F(t, x), \quad t \in [a, b], \quad x(a) \in C_0.$$

Here $x(t) \in \mathfrak{R}^n$, C is a closed set in $\mathfrak{R}^n$, and $F(t, x)$ is a multifunction on $[a, b] \times \mathfrak{R}^n$ with values in $\mathfrak{R}^n$.

Because of the obvious comparison with Clarke's [3] results, we use a similar setting. The following are the usual definitions for choosing appropriate multifunctions.

DEFINITION 5.1. A multifunction $\Gamma : \mathfrak{R}^m \to \mathfrak{R}^n$ is measurable, if for every open set C of $\mathfrak{R}^n$, the set

$$\{x \in \mathfrak{R}^m : \Gamma(x) \cap C \neq \varnothing\}$$

is Lebesgue measurable.

DEFINITION 5.2. A multifunction $F : [a, b] \times \mathfrak{R}^n \to 2^{\mathfrak{R}^n}$ is measurably Lipschitz if

(a) For each x in $\mathfrak{R}^n$, the multifunction $t \to F(t, x)$ is measurable on $[a, b]$.

(b) There is an integrable function $k(t)$ on $[a, b]$ such that for each t in $[a, b]$, the multifunction $x \to F(t, x)$ is nonempty and Lipschitz of rank $k(t)$.

For the rest of this section the interval $[a, b]$ will remain fixed and $\|x\|$ will denote the infinity norm:

$$\|x\| = \max \{|x(t)| : a \leqq t \leqq b\}.$$

The main result is the following.

THEOREM 5.3. *Assume that the following hypotheses hold*:

(i) *F is measurably Lipschitz, integrably bounded and closed on $[a, b] \times \mathfrak{R}^n$, and*

(ii) *f is Lipschitz of rank K_f.*

If x is an optimal solution of (P_2), then there exist an absolutely continuous arc p and constants K_1 and K_2 such that

$$[\dot{p}, p] \in \partial_B K_2 \rho(t, x, \dot{x}) \quad \text{a.e.}$$

$$p(a) \in \partial_B K_1 d_{C_0}(x(a)) \quad \text{and} \quad p(b) \in \partial_B f(x(b))$$

where ρ is as defined below.

The proof of this result takes up most of the remainder of this section.

To apply Proposition 1.1 a concept of distance from an arc x to the set of feasible arcs is required. The idea used here is based on the function $\rho : \mathfrak{R} \times \mathfrak{R}^n \times \mathfrak{R}^n \to [0, \infty]$ given by

$$\rho(t, x, v) = \inf \{|v - y| : y \in F(t, x)\}.$$

This distance relates to the distances in the coordinates as follows.

PROPOSITION 5.4 [3, Prop. 3.1.5]. *If F is measurably Lipschitz, then*

(a) *For each x and v in $\mathfrak{R}^n$, the function $t \to \rho(t, x, v)$ is measurable.*

(b) *For any (t, x_1) and (t, x_2) in $\mathfrak{R} \times \mathfrak{R}^n$, and for any v_1 and v_2 in $\mathfrak{R}^n$, we have*

$$|\rho(t, x_1, v_1) - \rho(t, x_2, v_2)| \leqq k(t)|x_1 - x_2| + |v_1 - v_2|.$$

A curve x is called a *trajectory* for F if

$$\dot{x}(t) \in F(t, x(t)) \quad \text{a.e.}$$

Associate a distance $\rho_F(x)$ for an arc x to be

$$\rho_F(x) = \int_a^b \rho(t, x(t), \dot{x}(t))\, dt.$$

The following theorem allows us to find a trajectory for F near a given arc x.

THEOREM 5.5 [3, Thm. 3.1.6]. *If x is an arc and $\rho_F(x) < \varepsilon/K$, then there is a trajectory y for F with $y(a) = x(a)$, $|x(t) - y(t)| < \varepsilon$ for all $t \in [a, b]$ and*

$$\|x - y\| \leqq \int_a^b |\dot{x}(t) - \dot{y}(t)|\, dt \leqq K\rho_F(x).$$

Here $K = \exp\{\int_a^b k(t)\, dt\}$.

LEMMA 5.6. *Let x be an optimal trajectory for the differential inclusion problem. Then, for some K_1 and K_2,*

$$\begin{aligned}\phi(y) &= f(y(b)) - f(x(b)) + K_1 d_{C_0}(y(a)) + K_2 \int_a^b \rho(t, y, \dot{y})\, dt \\ &\geqq 0\end{aligned} \tag{5.1}$$

for all trajectories y.

Proof. Suppose this is not true. Then there is an arc y with $\phi(y) < 0$.

Let $c \in C_0$ be such that $d_{C_0}(y(a)) = |y(a) - c|$ and define w by $w(t) = y(t) + c - y(a)$. By Proposition 5.4

$$\int_a^b \rho(t, w, \dot{w})\, dt \leqq \int_a^b \rho(t, y, \dot{y})\, dt + |y(a) - c| \ln(K).$$

From Theorem 5.5, there is an arc z for F such that $z(a) = w(a) = c$ and

$$\int_a^b |\dot{z} - \dot{w}|\, dt \leqq K \int_a^b \rho(t, w, \dot{w})\, dt.$$

Combining the above gives

$$\begin{aligned}\|z - y\| &\leqq \|z - w\| + \|w - y\| \\ &\leqq \int_a^b |\dot{z} - \dot{w}|\, dt + d_{C_0}(y(a)) \\ &\leqq K\left\{ d_{C_0}(y(a)) \ln(K) + \int_a^b \rho(t, y, \dot{y})\, dt \right\} + d_{C_0}(y(a)).\end{aligned}$$

Thus

$$\begin{aligned}\phi(z) &\leqq \phi(y) + K_f \|z - y\| \\ &\leqq \phi(y) + K_f K \int_a^b \rho(t, y, \dot{y})\, dt + K_f(K \ln(K) + 1) d_{C_0}(y(a)) \\ &< 0.\end{aligned} \tag{5.2}$$

This contradicts the optimality of x with $K_1 = K_f(K \ln(K) + 1)$ and $K_2 = K_f K$, and completes the proof of the lemma. □

To complete the proof of Theorem 5.3, we calculate the B-gradients of the functions $f(y(b)) - f(x(b))$, $K_1 d_{C_0}(y(a))$ and $K_2 \int_a^b \rho(t, y, \dot{y})\, dt$.

If ν is an element of the B-gradient of $f(y(b))-f(x(b))$ at x, then there is a $\psi_1 \in \partial_B f(x(b))$ such that

$$\nu(y) = \langle \psi_1, y(b) \rangle.$$

This follows from the chain rule result in [10].

Similarly, if $\nu \in \partial_B K_1 d_{C_0}(x(a))$, as a function of x, there is a $\psi_2 \in \partial_B K_1 d_{C_0}(x(a))$, as a function of a, such that

$$\nu(y) = \langle \psi_2, y(a) \rangle.$$

In a manner similar to that used in the previous section, an element ν in $\partial_B K_2 \int_a^b \rho(t, y, \dot{y})\, dt$ takes on the form

$$\nu(y) = \int_a^b [\langle q, y \rangle + \langle s, \dot{y} \rangle]\, dt$$

for some functions q and s with

$$(q, s) \in \partial_B K_2 \rho(t, y, \dot{y}) \quad \text{a.e.}$$

Combining these last three expressions with equation (5.1) shows that, for some ψ_1 and ψ_2 in $\partial_B f(x(b))$ and $\partial_B K_1 d_{C_0}(x(a))$ and a selection (q, s) of $\partial_B K_2 \rho(t, x, \dot{x})$,

$$0 = \langle \psi_1, y(a) \rangle + \langle \psi_2, y(b) \rangle + \int_a^b [\langle q, y \rangle + \langle s, \dot{y} \rangle]\, dt.$$

By the standard variational argument,

$$s(t) = \psi_1 + \int_a^t q\, d\tau, \quad \text{and} \quad s(b) = \psi_2.$$

If $p(t) = \psi_1 + \int_a^t q\, d\tau$, then p satisfies

$$[\dot{p}, p] \in \partial_B K_2 \rho(t, x, \dot{x}) \quad \text{a.e.}$$

$$p(a) \in \partial_B K_1 d_{C_0}(x(a)) \quad \text{and} \quad p(b) \in \partial_B f(x(b)).$$

Notes. (1) The condition that

$$[\dot{p}, p] \in \partial_B K_2 \rho(t, x, \dot{x})$$

translates into a well-known type of condition if $F(t, x)$ does not depend on x. Under this assumption, the condition implies the existence of an absolutely continuous function $p(t)$ such that

$$p(t) \in N_B(F(t), \dot{x}(t)) \quad \text{a.e.}$$

(2) The method used for stating the conclusions in Theorem 5.3 when using Clarke gradients involves the assumption that F is convex valued and the "true Hamiltonian,"

$$H(x, p) = \sup_y \{p \cdot y - \rho(x, y)\}.$$

Unfortunately, the current proofs relating the conditions in Theorem 5.3 to this Hamiltonian involve the semicontinuity properties of the Clarke gradients.

REFERENCES

[1] F. H. CLARKE, *Necessary conditions for nonsmooth problems in optimal control and the calculus of variations*, Ph.D. thesis, University of Washington, Seattle, WA, 1973.

[2] ———, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247-262.

[3] ———, *Optimization and Nonsmooth Analysis*, Wiley Interscience, New York, 1983.

[4] A. JOFRE AND L. THIBAULT, *Proximal and Frechet normal formulae for some small normal cones in Hilbert space*, to appear.

[5] A. D. IOFFE AND V. L. LEVIN, *Subdifferentials of convex function*, Trans. Moscow Math. Soc., 26 (1972), pp. 1-27.

[6] P. MICHEL AND J. P. PENOT, *Cacul sous-differential pour des functions Lipschitziennes et non-Lipschitziennes*, C. R. Acad. Sci. Paris, Ser. I Math., 298 (1985), pp. 269-272.

[7] J. P. PENOT AND P. TERPOLILLI, *Tangent cones and singularities*, C. R. Acad. Sci. Paris, Ser. I Math., 296 (1983), pp. 721-724.

[8] R. T. ROCKAFELLAR, *Generalized directional derivatives and subgradients of nonconvex functions*, Canad. J. Math., 32 (1980), pp. 157-180.

[9] J. S. TREIMAN, *Shrinking generalized gradients*, Nonlinear Anal., 12 (1988), pp. 1429-1449.

[10] ———, *Finite dimensional optimality conditions*: B-*gradients*, J. Optim. Theory Appl., 62 (1989), pp. 139-150.

[11] D. WARD, *Convex subcones of the contingent cone in nonsmooth calculus and optimization*, Trans. Amer. Math. Soc., 302 (1987), pp. 661-682.

[12] D. E. WARD AND R. MERKOVSKY, *Constraint qualifications in nondifferentiable optimization*, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 3, pp. 733-748, May 1990

015

EXACT CONTROLLABILITY OF THE ONE-DIMENSIONAL WAVE EQUATION WITH LOCALLY DISTRIBUTED CONTROL*

LOP FAT HO†

Abstract. The one-dimensional wave equation with variable wave speed and locally distributed control is considered. It is shown that the adjoint system is observable using a multiplier method with a multiplier being the solution of an ordinary differential equation. It is also shown that a sufficient condition for exact controllability is that a related minimization problem always has an optimal solution. Since the objective function for this minimization problem would be coercive if the adjoint system is observable this establishes the exact controllability of the original system.

Key words. wave equation, exact controllability, locally distributed control, adjoint system

AMS(MOS) subject classification. 93B05

1. Introduction. In this paper we consider the following control problem associated with the one-dimensional wave equation:

$$\rho(x)\frac{\partial^2}{\partial t^2}z(x,t)=\frac{\partial}{\partial x}\left(\sigma(x)\frac{\partial z}{\partial x}\right)+u(x,t), \qquad 0\leqq x\leqq l,\ t>0,$$

$$z(x,0)=\frac{\partial z}{\partial t}(x,0)=0, \qquad 0\leqq x\leqq l, \tag{1}$$

$$z(0,t)=z(l,t)=0, \qquad t>0,$$

with the additional assumption that

$$u(x,t)=0 \quad \text{for } x\notin[x_1,x_2] \tag{2}$$

where $0\leqq x_1<x_2\leqq l$. We assume that σ and ρ are C^1 and positive. Denote

$$E_1=H_0^1[0,l]\times L^2[0,l], \qquad U_1=L^2([0,l]\times[0,T]),$$

and

$$E_2=L^2[0,l]\times H^{-1}[0,l], \qquad U_2=L^2(0,T;H^{-1}[0,l]).$$

It has been shown in [3] that $u\in U_i$ implies that $(z,\partial z/\partial t)\in C(0,T;E_i)$ for any $T>0$, $i=1,2$.

The main result of this paper is the following theorem.

THEOREM 1. *Let $x_1, x_2\in[0,l]$, $x_1<x_2$, be fixed. Then for $i=1,2$, there exists $T_i^*>0$ such that given any pair $(\phi^0,\phi^1)\in E_i$, there exists $u\in U_i$, u vanishing outside $[x_1,x_2]\times[0,T]$, such that the solution of (1) satisfies*

$$z(x,T)=\phi^0(x,T) \quad \text{and} \quad \frac{\partial z(x,T)}{\partial t}=\phi^1(x,T), \quad 0\leqq x\leqq l.$$

(*The equalities are to be interpreted in the sense of distributions.*)

It is known that the controllability of a control system is related to the observability of its adjoint system [1]. Thus the controllability problem can be reduced to an inequality for the adjoint system (see [4]). For the wave equation with control in the Dirichlet boundary condition, such an inequality was proved using a multiplier method in [2].

* Received by the editors May 15, 1989; accepted for publication August 8, 1989.

† Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.

In § 2, we will state an observability result. Three observation operators are claimed to be boundedly invertible. In § 3, we prove the easiest part, the bounded invertibility of the first operator. We also prove that the remaining operators are injective. The proof for the bounded invertibility is based on a multiplier technique. The multiplier we use is the solution of an ordinary differential equation with parameters that depend on the functions ρ, σ, and the interval $[x_1, x_2]$. In § 4, we prove the more difficult parts, the bounded invertibility of the remaining two operators. We need to further apply the multiplier method and use a uniqueness and compactness argument. Finally, in § 5, we prove Theorem 1, the controllability result from Theorem 2, the observability result. The method of proof is new. It is based on an observation that a sufficient condition for the control problem to have a solution is that a related minimization problem has an optimal solution. This relationship between a control problem and a minimization problem holds for many systems, in particular, for any system for which the method in [4] can be applied.

2. Observability. It is well known that controllability results follow from observability of the adjoint system [1]. In a rather general context, we can say that observability implies exact controllability to *some* space. (See the Hilbert Uniqueness Method (HUM) in [4].) However, it may be difficult, in some cases, to characterize this space in a concrete manner.

Associated with the system (1), we will consider its adjoint system

$$\rho \frac{\partial^2 y}{\partial t^2}(x, t) = \frac{\partial}{\partial x}\left(\sigma(x)\frac{\partial y}{\partial x}(x, t)\right), \qquad 0 \leqq x \leqq l, \quad t > 0,$$

$$y(0, t) = y(l, t) = 0, \qquad t > 0, \tag{3}$$

$$y(x, 0) = \psi_0(x), \quad \frac{\partial y}{\partial t}(x, 0) = \psi_1(x), \quad \psi_0 \in H_0^1[0, l], \quad \psi_1 \in L^2[0, l].$$

Let $T_1, T_2, T_3 > 0$ and $0 \leqq x_1 \leqq x_2 \leqq l$ be given. We will look at three different observation operators:

(i) $$\mathscr{H}_1 : H_0^1[0, l] \times L^2[0, l] \to L^2([x_1, x_2] \times [0, T_1])^2,$$

$$\mathscr{H}_1(\psi_0, \psi_1) = \left(\frac{\partial y}{\partial t}, \frac{\partial y}{\partial x}\right)\Bigg|_{[x_1, x_2] \times [0, T_1]};$$

(ii) $$\mathscr{H}_2 : H_0^1[0, l] \times L^2[0, l] \to L^2([x_1, x_2] \times [0, T_2]),$$

$$\mathscr{H}_2(\psi_0, \psi_1) = \frac{\partial y}{\partial t}\Bigg|_{[x_1, x_2] \times [0, T_2]};$$

(iii) $$\mathscr{H}_3 : H_0^1[0, l] \times L^2[0, l] \to L^2([x_1, x_2] \times [0, T_3]),$$

$$\mathscr{H}_3(\psi_0, \psi_1) = \frac{\partial y}{\partial x}\Bigg|_{[x_1, x_2] \times [0, T_3]}.$$

THEOREM 2. *There exist $T_1, T_2, T_3 > 0$ such that the operators $\mathscr{H}_1, \mathscr{H}_2, \mathscr{H}_3$ are boundedly invertible.*

Actually, we will find, for $i = 1, 2, 3$, $T_i^* > 0$ such that $T_i > T_i^*$ implies that $\mathscr{H}_i$ is boundedly invertible, $i = 1, 2, 3$.

Stated in a more precise way, Theorem 1 claims that there exist constants $K_1, K_2, K_3 > 0$ such that for all solutions y of (3), the following inequalities hold:

$$\frac{1}{2}\int_0^T \int_{x_1}^{x_2} \rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\,dt \geqq K_1 E_0, \tag{4}$$

$$\frac{1}{2}\int_0^T \int_{x_1}^{x_2} \rho\left(\frac{\partial y}{\partial t}\right)^2 dx\,dt \geqq K_2 E_0, \tag{5}$$

$$\frac{1}{2}\int_0^T \int_{x_2}^{x_1} \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\,dt \geqq K_3 E_0, \tag{6}$$

where

$$E_0 = \frac{1}{2}\int_0^l \rho\left(\frac{\partial y}{\partial t}(x,0)\right)^2 + \sigma\left(\frac{\partial y}{\partial x}(x,0)\right)^2 dx. \tag{7}$$

By Poincaré's inequality, we know that the norm

$$\left(\frac{1}{2}\int_0^l f'\,dx\right)^{1/2}$$

of f is equivalent with the H_0^1 norm of f. So inequalities (4), (5), and (6) together are indeed equivalent to Theorem 2.

Clearly, both (5) and (6) are stronger than (4). When the wave speed c is a constant, the first inequality (4) is not difficult to prove because in that case, we can solve y explicitly in terms of the initial functions ψ_0 and ψ_1. For T_1 sufficiently large, we can then go back and solve ψ_0 and ψ_1 in terms of the values of y on $[x_1, x_2]\times[0, T_1]$. However, when c is not a constant, there are no explicit formulas for y, and the proof of this inequality is more difficult.

3. Bounded invertibility of the operator $\mathscr{H}_1$ and weak observability results. We start with the easier result, namely, the bounded invertibility of $\mathscr{H}_1$. In fact, we will prove the inequality

$$\frac{1}{2}\int_0^T \int_{x_2}^{x_1} \rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial t}\right)^2 dx\,dt \geqq K(T - T_1^*)^+ E_0 \tag{8}$$

where K and T_1^* are positive constants.

Proof of (8). Let g be any real-value continuous and piecewise C^1 function defined on $[0, l]$ such that $g(0) = g(l) = 0$. On multiplying the equality $\rho(\partial^2 y/\partial t^2) - \partial/\partial x(\sigma(\partial y/\partial x)) = 0$ on both sides by $(\partial y/\partial x)$ and integrating, we obtain

$$\begin{aligned} 0 &= \int_0^l g\frac{\partial y}{\partial x}\left(\rho\frac{\partial^2 y}{\partial t^2} - \frac{\partial}{\partial x}\left(\sigma\frac{\partial y}{\partial x}\right)\right) dx \\ &= -\frac{1}{2}\int_0^l g\left(\rho\frac{\partial}{\partial x}\left(\frac{\partial y}{\partial t}\right)^2 + \frac{1}{\sigma}\frac{\partial}{\partial x}\left(\sigma\frac{\partial y}{\partial x}\right)^2\right) dx + \frac{\partial}{\partial t}\int_0^l g\rho\frac{\partial y}{\partial x}\frac{\partial y}{\partial t}\,dx \\ &= \frac{1}{2}\int_0^l \left(\frac{d}{dx}g\rho\right)\left(\frac{\partial y}{\partial t}\right)^2 + \frac{d}{dx}\left(\frac{g}{\sigma}\right)\left(\sigma\frac{\partial y}{\partial x}\right)^2 dx + \frac{\partial}{\partial t}\int_0^l g\rho\frac{\partial y}{\partial x}\frac{\partial y}{\partial t}\,dx. \end{aligned} \tag{9}$$

For $\lambda \in [0, 1]$, let g_λ be the continuous and piecewise C^1 solution of the differential

equation

$$g'_\lambda = \begin{cases} \lambda + \min\left\{\dfrac{\sigma'}{\sigma} g_\lambda, -\dfrac{\rho'}{\rho} g_\lambda\right\} & \text{on } [x_1, x_2], \\ \lambda - 1 + \min\left\{\dfrac{\sigma'}{\sigma} g_\lambda, -\dfrac{\rho'}{\rho} g_\lambda\right\} & \text{on } [0, l]\backslash[x_1, x_2], \end{cases}$$

$$g_\lambda(0) = 0.$$

For $\lambda = 0$, we clearly have $g_0 \leqq 0$ on $[0, l]$. Hence $g_0(l) \leqq 0$. For $\lambda = 1$, $g_1 \geqq 0$ on $[0, l]$ and $g'_1 > 0$ on $[x_1, x_2]$. So $g_1(l) > 0$. But $g_\lambda(l)$ depends continuously on λ. So there exists $\lambda \in [0, 1)$ such that $g_\lambda(l) = 0$.

On (x_1, x_2),

$$\frac{d}{dx}(g_\lambda \rho) = g'_\lambda \rho + g_\lambda \rho' \leqq \lambda \rho$$

and

$$\frac{d}{dx}\left(\frac{g_\lambda}{\sigma}\right) = \frac{g'_\lambda}{\sigma} - \frac{\sigma' g_\lambda}{\sigma^2} \leqq \frac{\lambda}{\sigma}.$$

On $[0, l]\backslash[x_1, x_2]$,

$$\frac{d}{dx}(g_\lambda \rho) \leqq (\lambda - 1)\rho \quad \text{and} \quad \frac{d}{dx}\left(\frac{g_\lambda}{\sigma}\right) \leqq \frac{\lambda - 1}{\sigma}.$$

Hence by (9), we obtain

$$0 \leqq \int_{x_2}^{x_1} \lambda\left(\rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2\right) dx + \int_{[0,x_1]\cup[x_2,l]} (\lambda - 1)\left(\rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2\right) dx + \frac{\partial}{\partial t}\int_0^l g_\lambda \rho \frac{\partial y}{\partial x}\frac{\partial y}{\partial t}\, dx.$$

Therefore,

$$\text{(10)} \quad \int_{x_2}^{x_1} \rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx \geqq (1 - \lambda)\int_0^l \rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx - \frac{\partial}{\partial t}\int_0^l g_\lambda \rho \frac{\partial y}{\partial x}\frac{\partial y}{\partial t}\, dx.$$

Note that for any t,

$$\left|\int_0^l g_\lambda \rho \frac{\partial y}{\partial x}\frac{\partial y}{\partial t}\, dx\right| \leqq \frac{M}{2}\int_0^l \rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx$$

$$= ME_0 \text{ (by conservation of energy).}$$

Here,

$$M = \max\{g_\lambda(x)(\rho(x)/\sigma(x))^{1/2} : 0 \leqq x \leqq l\}.$$

Hence, integrating (10) from $t = 0$ to $t = T$, we have

$$\int_0^T \int_{x_2}^{x_1} \rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\, dt$$

$$\geqq (1 - \lambda)\int_0^T \int_0^l \rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\, dt - 2ME_0$$

$$= ((1 - \lambda)T - 2M)E_0 \quad \text{(by conservation of energy).}$$

So (8) holds with $K = 1 - \lambda$ and $T_1^* = 2M/(1 - \lambda)$. □

Remark. When $c(x)=(\sigma(x)/\rho(x))^{1/2}$ is constant, it is not hard to see that $1-\lambda(x_2-x_1)/l$ and $M=\max\{(x_2-x_1)x_1,\ (x_2-x_1)(l-x_2)\}/(cl)$. Hence

$$T_1^*=\frac{2}{c}\max\{x_1,\ l-x_2\}$$

$$=\frac{2}{c}\times \text{distance from the set } [x_1,\ x_2] \text{ to the boundary points } \{0,\ l\}.$$

In this case, we have a simple physical interpretation for T_1^*. Since waves travel with speed c on the string, T_1^* is the largest length of any period of time during which a signal can stay entirely outside the interval $[x_1, x_2]$.

Also, in this case (with $c=$ constant), $K=(x_2-x_1)/l$. So inequality (10) is equivalent to

$$\frac{1}{2(x_1-x_2)}\int_0^T\int_{x_2}^{x_1}\rho\left(\frac{\partial y}{\partial t}\right)^2+\sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\,dt\geqq\frac{1}{l}\left(T-\frac{2}{c}\max\{x_1,\ l-x_2\}\right)E_0.$$

Letting $x_2\to x_1$, we see that the inequality

$$\frac{1}{2}\int_0^T\rho\left(\frac{\partial y}{\partial t}\right)^2+\sigma\left(\frac{\partial y}{\partial x}\right)^2 dt\geqq\frac{1}{l}\left(T-\frac{2}{c}\max\{x_1,\ l-x_1\}\right)E_0 \tag{11}$$

holds, giving us another observability result. In fact, we can show that the above inequality holds even for variable c, with, of course, different constants.

However, we can easily give counterexamples (for some x) such that neither of the inequalities

$$\frac{1}{2}\int_0^T\rho\left(\frac{\partial y}{\partial t}(x,t)\right)^2 dt\geqq K_2E_0 \tag{12}$$

nor

$$\frac{1}{2}\int_0^T\sigma\left(\frac{\partial y}{\partial x}(x,t)\right)^2 dt\geqq K_3E_0 \tag{13}$$

holds for any $T>0$, no matter how large. Actually, we can show that these inequalities cannot hold for *any* $x\in[0,\ l]$. The validity of (12) and (13) is related to pointwise control problems (see [4]).

Next, we will establish two weak observability results. These results prepare the way for the stronger inequalities (5), (6), which we will prove in the next section.

PROPOSITION 3. *Let T_1^* be such that* (8) *holds. Then for $T_2>T_1^*$, the operator $\mathscr{H}_2$ is one to one.*

Proof. We first note that in the proof of (8), the boundary conditions $y(0,t)=y(l,t)=0$ are only used to guarantee that energy is conserved. Since energy is also conserved for any boundary condition $y(0,t)=\alpha$, $y(l,t)=\beta$, α, β being constants, the same inequality (8) holds for solutions of the wave equation satisfying these boundary conditions.

Let us now suppose that for some pair of initial functions (ψ_0,ψ_1), we have $\mathscr{H}_2((\psi_0,\psi_1))=0$. This means that the solution y of (3) satisfies

$$\frac{\partial y}{\partial t}=0\quad\text{on } [x_1,x_2]\times[0,T_2],$$

where $T_2 > T_1^*$. It follows that $\partial/\partial x(\sigma(\partial y/\partial x)) = \rho(\partial^2 y/\partial t^2) = 0$ on $[x_1, x_2]\times[0, T_2]$. Hence $y(x, t) = c_1(t) + c_2(t)p(x)$ there, where $p'(x) = 1/\sigma(x)$. Because $\partial y/\partial t = 0$, c_1, c_2 must be constants. Then defining

$$y_1(x, t) = y(x, t) - c_1 - c_2 p(x),$$

we see that y_1 is also a solution of the wave equation with boundary conditions

$$y_1(0, t) = -c_1, \qquad y_1(l, t) = -c_1 - c_2 p(l).$$

So by the remark we have just made, (8) holds for $y = y_1$. Hence

$$\frac{1}{2}\int_0^{T_2}\int_{x_1}^{x_2} \rho\left(\frac{\partial y_1}{\partial t}\right)^2 + \sigma\left(\frac{\partial y_1}{\partial x}\right)^2 dx\,dt$$

$$\geqq K(T_2 - T_1^*)^+ \frac{1}{2}\int_0^l \rho\left(\frac{\partial y_1(x, 0)}{\partial t}\right)^2 + \sigma\left(\frac{\partial y_1(x, 0)}{\partial x}\right)^2 dx.$$

Because the left-hand side vanishes, we must have

$$\frac{\partial y_1(x, 0)}{\partial t} = \frac{\partial y_1(x, 0)}{\partial x} = 0 \quad \text{on } [0, l].$$

Hence

$$\psi_1(x) = \frac{\partial y(x, 0)}{\partial t} = \frac{\partial y_1(x, 0)}{\partial t} = 0 \quad \text{on } [0, l]$$

and

$$\frac{\partial y(x, 0)}{\partial x} = \frac{\partial y_1(x, 0)}{\partial x} - \frac{c_2}{\sigma(x)} = -\frac{c_2}{\sigma(x)}.$$

So $\psi_0(x) = y(x, 0) = -c_2 p(x) + c_3$, c_3 a constant. But $y(0, 0) = y(1, 0) = 0$. This implies that $c_2 = c_3 = 0$. Hence $\psi_0(x) = 0$ on $[0, l]$. So we have $(\psi_0, \psi_1) = (0, 0)$. It follows that $\mathcal{H}_2$ is one to one. □

PROPOSITION 4. *There exists T_3^* such that $\mathcal{H}_3$ is one to one whenever $T_3 \geqq T_3^*$.*

Proof. Let $\mathcal{H}_3((\psi_0, \psi_1)) = 0$. Then the solution y of (3) satisfies

$$\frac{\partial y}{\partial x} = 0 \quad \text{on } [x_1, x_2]\times[0, T_3].$$

Hence

$$\frac{\partial^2 y}{\partial t^2} = \frac{1}{\rho}\frac{\partial}{\partial x}\left(\sigma\frac{\partial y}{\partial x}\right) = 0.$$

Therefore

$$y(x, t) = c_1(x)t + c_2(x) \quad \text{on } [x_1, x_2]\times[0, T_3].$$

Because $\partial y/\partial x = 0$ on $[x_1, x_2]\times[0, T_3]$, it follows that c_1, c_2 are constants. By (8), we then have

$$K(T_3 - T_1^*)^+ E_0 \leqq \frac{1}{2}\int_0^{T_3}\int_{x_1}^{x_2} \rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\,dt$$

(14)

$$\leqq \frac{m}{2} T_3(x_2 - x_1)c_1^2$$

where $m = \sup\{\rho(x): 0 \leqq x \leqq l\}$. On the other hand, applying Poincaré's inequality and

using conservation of energy again, we have

$$
\begin{aligned}
E_0 &= \frac{1}{2}\int_0^l \rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx \\
&\geqq \frac{1}{2} m_1 \left(\frac{\pi}{l}\right)^2 \int_0^l y^2\, dx \\
&\geqq \frac{1}{2} m_1 \left(\frac{\pi}{l}\right)^2 \int_{x_1}^{x_2} y^2\, dx \\
&= \frac{1}{2} m_1 \left(\frac{\pi}{l}\right)^2 (c_1 t + c_2)^2 (x_2 - x_1)
\end{aligned} \tag{15}
$$

for $0 \leqq t \leqq T_3$, where $m_1 = \min\{|\sigma(x)|: x \in [0, l]\}$. Eliminating E_0 from (14) and (15), we obtain (we may assume $T_3 > T_1^*$)

$$
(c_1 t + c_2)^2 \leqq \left(\frac{l}{\pi}\right)^2 \frac{m T_3 c_1^2}{k m_1 (T_3 - T_1^*)}
$$

for all $0 \leqq t \leqq T_3$. Note that

$$
\min_{c_2} \max_{0 \leqq t \leqq T_3} (c_1 t + c_2)^2 = \left(\frac{c_1 T_3}{2}\right)^2.
$$

So

$$
\left(\frac{c_1 T_3}{2}\right)^2 \leqq \left(\frac{l}{\pi}\right)^2 \frac{m T_3 c_1^2}{k m_1 (T_3 - T_1^*)}.
$$

Hence

$$
c_1^2 T_3 \left(T_3^2 - T_1^* T_3 - \left(\frac{l}{\pi}\right)^2 \frac{4m}{m_1 k} \right) \leqq 0. \tag{16}
$$

Let $T_3^* = \frac{1}{2}(T_1^* + \sqrt{T_1^{*2} + (16 m l^2 / k m_1 \pi^2)})$. Then $T_3^* > T_3$ implies

$$
T_3^2 - T_1^* T_3 - \left(\frac{l}{\pi}\right)^2 \frac{4m}{k m_1} > 0. \tag{17}
$$

From (16) and (17), we see that we must have $c_1 = 0$. So $\partial y / \partial t = 0$ on $[x_1, x_2] \times [0, T]$. Since $T_3 > T_1^*$, by the previous proposition, $\psi_0 = \psi_1 = 0$. Hence we have proved that $\mathscr{H}_3$ is one to one. □

4. Bounded invertibility of $\mathscr{H}_2$ and $\mathscr{H}_3$. We will now prove the following more difficult inequalities:

$$
\frac{1}{2}\int_0^T \int_{x_1}^{x_2} \rho\left(\frac{\partial y}{\partial t}\right)^2 dx\, dt \geqq K_2 E_0, \tag{18}
$$

$$
\frac{1}{2}\int_0^T \int_{x_2}^{x_1} \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\, dt \leqq K_3 E_0. \tag{19}
$$

We note that the left-hand side of (8) is the total energy of y on $[x_1, x_2] \times [0, T]$, whereas that of (18) and (19) are, respectively, the total kinetic energy and the total potential energy of y on $[x_1, x_2] \times [0, T]$. So if we are able to show that the difference between the total kinetic energy and the total potential energy on $[x_1, x_2] \times [0, T]$ is "small" in an appropriate sense, then (18) and (19) will follow from (8). This is, in fact, true when $x_1 = 0$ and $x_2 = l$ and is what is known as *equipartition of energy*:

$$
\left| \frac{1}{2}\int_0^T \int_0^l \rho\left(\frac{\partial y}{\partial t}\right)^2 - \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\, dt \right|^2 \leqq K E_0,
$$

where K is a constant independent of y and T. Unfortunately, we are not able to prove such an inequality when 0 and l are replaced, respectively, by arbitrary x_1 and x_2. However, we will borrow some ideas from the method of proof of the equipartition of energy. The first step of our proof is the following lemma.

LEMMA 5. *There exist constants K_1 and K_2 such that for any $x_0 \in [0, 1]$, any $T > 0$, and any solution y of (3), we have*

$$\frac{1}{2}\int_0^T \rho(x_0)\left(\frac{\partial y(x_0,t)}{\partial t}\right)^2 + \sigma(x_0)\left(\frac{\partial y(x_0,t)}{\partial x}\right)^2 dt \leqq (K_1 T + K_2)E_0. \tag{20}$$

Proof. We multiply the equation

$$\rho\frac{\partial^2 y}{\partial t^2} - \frac{\partial}{\partial x}\left(\sigma\frac{\partial y}{\partial x}\right) = 0 \tag{21}$$

on both sides by $x(\partial y/\partial x)$. Integrating from $x = 0$ to $x = x_0$, rearranging, and integrating by parts, we obtain

$$\begin{aligned}&\frac{x_0}{2}\left(\rho(x_0)\left(\frac{\partial y(x_0,t)}{\partial t}\right)^2 + \sigma(x_0)\left(\frac{\partial y(x_0,t)}{\partial x}\right)^2\right)\\ &\quad = \frac{1}{2}\int_0^{x_0}\frac{d}{dx}(x\rho)\left(\frac{\partial y}{\partial t}\right)^2 + \frac{d}{dx}\left(\frac{x}{\sigma}\right)\left(\sigma\frac{\partial y}{\partial x}\right)^2 dx + \frac{\partial}{\partial t}\int_0^{x_0} x\rho\frac{\partial y}{\partial x}\frac{\partial y}{\partial t}\,dx.\end{aligned} \tag{22}$$

Similarly, multiplying (20) by $(l-x)\partial y/\partial x$, integrating from $x = x_0$ to $x = l$, and rearranging, we obtain

$$\begin{aligned}&\frac{1}{2}(l-x_0)\left(\rho(x_0)\left(\frac{\partial y(x_0,t)}{\partial t}\right)^2 + \sigma(x_0)\left(\frac{\partial y(x_0,t)}{\partial x}\right)^2\right)\\ &\quad = \frac{1}{2}\int_{x_0}^{l}\frac{d}{dx}\left((l-x)\rho\left(\frac{\partial y}{\partial t}\right)^2 + \frac{d}{dx}\left(\frac{l-x}{\sigma}\right)\left(\sigma\frac{\partial y}{\partial x}\right)^2 dx\right.\\ &\qquad \left. + \frac{\partial}{\partial t}\int_0^{x_0}(l-x)\rho\frac{\partial y}{\partial x}\frac{\partial y}{\partial t}\,dx.\right.\end{aligned} \tag{23}$$

Hence, combining (22) and (23), we have

$$\begin{aligned}&\frac{l}{2}\left(\rho(x_0)\left(\frac{\partial y(x_0,t)}{\partial t}\right)^2 + \sigma(x_0)\left(\frac{\partial y(x_0,t)}{\partial x}\right)^2\right)\\ &\quad = \frac{1}{2}\int_0^l \rho h\left(\frac{\partial y}{\partial t}\right)^2 + \sigma f\left(\frac{\partial y}{\partial x}\right)^2 dx + \frac{\partial}{\partial t}\int_0^l r\rho\frac{\partial y}{\partial x}\frac{\partial y}{\partial t}\,dx\end{aligned} \tag{24}$$

where

$$h(x) = \begin{cases}\dfrac{1}{\rho}\dfrac{d}{dx}(x\rho) & \text{on } [0, x_0],\\[2ex] \dfrac{1}{\rho}\dfrac{d}{dx}(x-l)\rho & \text{on } [x_0, l],\end{cases}$$

$$f(x) = \begin{cases}\dfrac{1}{\sigma}\dfrac{d}{dx}\left(\dfrac{x}{\sigma}\right)m & \text{on } [0, x_0],\\[2ex] \dfrac{1}{\sigma}\dfrac{d}{dx}\left(\dfrac{x-l}{\sigma}\right) & \text{on } [x_0, l],\end{cases}$$

and

$$r(x)=\begin{cases} x & \text{on } [0, x_0], \\ l-x & \text{on } [x_0, l]. \end{cases}$$

Let

$$M_1=\max\{\max\{h(x), f(x)\}: 0\leqq x\leqq l\}$$

and

$$M_2=\max\left\{\frac{r(x)}{c(x)}: 0\leqq x\leqq l\right\}, \quad \text{where } c(x)=\left(\frac{\sigma(x)}{\rho(x)}\right)^{1/2}.$$

Then by integrating (24) from $t=0$ to $t=T$, we obtain

$$\begin{aligned}\frac{l}{2}\int_0^T \rho(x_0)\left(\frac{\partial y(x_0,t)}{\partial t}\right)^2+\sigma(x_0)\left(\frac{\partial y(x_0,t)}{\partial x}\right)^2 dt \\ \leqq \frac{M_1}{2}\int_0^T\int_0^l \rho\left(\frac{\partial y}{\partial t}\right)^2+\sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\,dt+2M_2E_0 \\ =\left(\frac{M_1T}{2}+2M_2\right)E_0.\end{aligned}$$

This proves (20) with $K_1=M_1/2l$ and $K_2=2M_2/l$. □

Next, we give an estimate for the difference between the kinetic energy and the potential energy, as we do when proving the equipartition of energy. We show that the integral of the difference is less than a small term involving the total energy and other terms involving the integral of y^2 (not involving the derivatives of y).

LEMMA 6. *Let y be a solution of* (3) *and let*

$$D=\frac{1}{2}\int_0^T\int_{x_1}^{x_2}\rho\left(\frac{\partial y}{\partial t}\right)^2-\sigma\left(\frac{\partial y}{\partial x}\right)^2 dx. \tag{25}$$

Then we have for any $\varepsilon>0$,

$$\begin{aligned}|D|\leqq M\Bigg(&\coth\left(\frac{x_2-x_1}{2\varepsilon^2}\right)\int_0^T\int_{x_1}^{x_2}\frac{1}{\varepsilon^3}y(x,t)^2\,dx\,dt \\ &+\frac{1}{2\varepsilon}\int_{x_1}^{x_2}y(x,0)^2+y(x,T)^2\,dx \\ &+\frac{\varepsilon}{2}\left(K_T+2T\coth\left(\frac{x_2-x_1}{2\varepsilon^2}\right)\right)E_0\Bigg)\end{aligned} \tag{26}$$

where K_T and M are constants independent of ε. (K_T may depend on T.)

Proof. Multiplying the equation

$$\rho\frac{\partial^2 y}{\partial t^2}-\frac{\partial}{\partial x}\left(\sigma\frac{\partial y}{\partial x}\right)=0$$

by y and integrating, we have

$$\begin{aligned}0&=\int_0^T\int_{x_1}^{x_2}y\left(\rho\frac{\partial^2 y}{\partial t^2}-\frac{\partial}{\partial x}\left(\sigma\frac{\partial y}{\partial x}\right)\right)dx\,dt \\ &=-\int_0^T\int_{x_1}^{x_2}\rho\left(\frac{\partial y}{\partial t}\right)^2-\sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\,dt+\int_{x_1}^{x_2}\rho y\frac{\partial y}{\partial t}\,dx\,\Bigg|_{t=0}^{t=T}-\int_0^T\sigma y\frac{\partial y}{\partial x}\,dt\,\Bigg|_{x=x_0}^{x=x_1}.\end{aligned}$$

Hence $D = I_1 + I_2$ where

$$I_1 = \int_{x_1}^{x_2} \rho y \frac{\partial y}{\partial t}\, dx \Bigg|_{t=0}^{t=T}$$

and

$$I_2 = -\int_0^T \sigma y \frac{\partial y}{\partial x}\, dt \Bigg|_{x=x_1}^{x=x_2}.$$

Let $\varepsilon > 0$. For any t, we have

$$\left| \int_{x_1}^{x_2} \rho y \frac{\partial y}{\partial t}\, dx \right| \leqq \frac{M_1}{2} \int_{x_1}^{x_2} \frac{y^2}{\varepsilon} + \varepsilon \rho \left(\frac{\partial y}{\partial t}\right)^2 dx$$

$$\leqq M_1 \left(\frac{1}{2\varepsilon} \int_{x_1}^{x_2} y^2\, dx + \varepsilon E_0 \right)$$

where

$$M_1 = \sup \{|\rho(x)^{1/2}|: 0 \leqq x \leqq l\}.$$

Hence

$$I_1 \leqq M_1 \left(2\varepsilon E_0 + \frac{1}{2\varepsilon} \int_{x_1}^{x_2} y(x, 0)^2 + y(x, T)^2\, dx \right). \tag{27}$$

Also,

$$I_2 \leqq \frac{1}{2} M_2 \int_0^T \varepsilon \left(\sigma(x_1) \left(\frac{\partial y(x_1, t)}{\partial x}\right)^2 + \sigma(x_2) \left(\frac{\partial y(x_2, t)}{\partial x}\right)^2 \right) + \frac{1}{\varepsilon} (y(x_1, t)^2 + y(x_2, t)^2)\, dt \tag{28}$$

where $M_2 = \sup \{\sigma(x)^{1/2}: 0 \leqq x \leqq l\}$. The first term on the right-hand side can be estimated using Lemma 5. For the second term, we note that for any $\varepsilon > 0$,

$$y(x_1, t)^2 + y(x_2, t)^2 \leqq \coth \left(\frac{x_2 - x_1}{2\varepsilon^2} \right) \int_{x_1}^{x_2} \frac{1}{\varepsilon^2} y(x, t)^2 + \varepsilon^2 \left(\frac{\partial y(x, t)}{\partial x} \right)^2 dx. \tag{29}$$

The above inequality follows from a special case of the Trace Theorem. The explicit constants can be obtained by the calculus of variations. Hence

$$\int_0^T y(x_1, t)^2 + y(x_2, t)^2\, dt$$

$$\leqq \coth \left(\frac{x_2 - x_1}{2\varepsilon^2} \right) \left(\int_0^T \int_{x_1}^{x_2} \frac{1}{\varepsilon^2} y(x, t)^2\, dx\, dt + \varepsilon^2 2TM_3 E_0 \right),$$

where $M_3 = \max \{\sigma(x)^{-1/2}: 0 \leqq x \leqq T\}$.

It follows from (28), Lemma 5, and the above inequality that

$$I_2 \leqq M_2 \left(\frac{\varepsilon}{2} (K_1 T + K_2 + 2TM_3 \coth \left(\frac{x_2 - x_1}{2\varepsilon^2} \right)) E_0 + \coth \left(\frac{x_2 - x_1}{2\varepsilon^2} \right) \int_0^T \int_{x_1}^{x_2} \frac{1}{\varepsilon^3} y(x, t)^2\, dx\, dt \right). \tag{30}$$

Combining (27) and (30), we obtain the desired inequality. □

We are now ready to prove (5) and (6). Recall that we have already proved

$$\frac{1}{2}\int_0^T \int_{x_1}^{x_2} \rho\left(\frac{\partial y}{\partial t}\right)^2 + \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\,dt \geqq K(T-T_1^*)^+ E_0. \tag{8}$$

It follows that we have

$$\int_0^T \int_{x_1}^{x_2} \rho\left(\frac{\partial y}{\partial t}\right)^2 dx\,dt \geqq K(T-T_1^*)^+ E_0 - |D| \tag{31}$$

and

$$\int_0^T \int_{x_1}^{x_2} \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\,dt \geqq K(T-T_1^*)^+ E_0 - |D| \tag{32}$$

where D is as defined in (25). Let $T > T_1^*$. Let $\varepsilon > 0$ be sufficiently small so that

$$M\varepsilon\left(K_T + 2T\coth\left(\frac{x_2-x_1}{2\varepsilon^2}\right)\right) \leqq K(T-T_1^*).$$

It follows from Lemma 6 that

$$|D| < \frac{1}{2}K(T-T_1^*)E_0 + c_1\int_{x_1}^{x_2} y(x,0)^2 + y(x,T)^2\,dx + c_2\int_0^T\int_{x_1}^{x_2} y(x,t)^2\,dx\,dt,$$

where $c_1 = M/2\varepsilon$ and $c_2 = M/\varepsilon^3 \coth((x_2-x_1)/2\varepsilon^2)$. Hence (31) and (32) imply, respectively, that

$$\begin{aligned}\int_0^T \int_{x_1}^{x_2} \rho\left(\frac{\partial y}{\partial t}\right)^2 dx\,dt + c_1\int_{x_1}^{x_2} y(x,0)^2 + y(x,T)^2\,dx \\ + c_2\int_0^T\int_{x_1}^{x_2} y(x,t)^2\,dx\,dt \geqq \frac{1}{2}K(T-T_1^*)E_0\end{aligned} \tag{33}$$

and

$$\begin{aligned}\int_0^T \int_{x_1}^{x_2} \sigma\left(\frac{\partial y}{\partial x}\right)^2 dx\,dt + c_1\int_{x_1}^{x_2} y(x,0)^2 + y(x,T)^2\,dx \\ + c_2\int_0^T\int_{x_1}^{x_2} y(x,t)^2\,dx\,dt \geqq \frac{1}{2}K(T-T_1^*)E_0.\end{aligned} \tag{34}$$

We introduce the operator

$$\mathcal{F}: E_1 = H_0^1[0,l]\times L^2[0,l] \to L^2([x_1,x_2]\times L^2[x_1,x_2]) \times L^2([x_1,x_2]\times[0,T])$$

on the space of initial states by

$$\mathcal{F}(\phi,\psi) = (\sqrt{c_1}\,y(\cdot,0), \sqrt{c_1}\,y(\cdot,t), \sqrt{c_2}\,y).$$

We claim that $\mathcal{F}$ is compact. Let $\mathcal{F}_1: E_1 \to L^2[x_1,x_2]$, $\mathcal{F}_2: E_1 \to L^2[x_1,x_2]$, and $\mathcal{F}_3: E_1 \to L^2([x_1,x_2]\times[0,T])$ be defined, respectively, by

$$\mathcal{F}_1(\phi,\psi) = y(\cdot,0),$$

$$\mathcal{F}_2(\phi,\psi) = y(\cdot,T),$$

$$\mathcal{F}_3(\phi,\psi) = y.$$

It suffices to show that $\mathscr{F}_i$, $i = 1, 2, 3$, are compact. Denote the imbedding from $H_0^1[0, l]$ into $L^2[0, l]$ by $\mathscr{I}$. Then $\mathscr{I}$ is compact. Because $\mathscr{F}_1 = \mathscr{I} \circ \mathscr{P}$ where $\mathscr{P}$ is the projection mapping (ϕ^0, ϕ^1) to ϕ^0, $\mathscr{F}_1$ is compact. Let $\mathscr{S}$ be the mapping carrying $(y(\cdot, 0), \partial y(\cdot, 0)/\partial x)$ to $(y(\cdot, T), \partial y(\cdot, T)/\partial x)$. By conservation of energy, $\mathscr{S}$ is an isomorphism from E_1 onto itself. Because $\mathscr{F}_2 = \mathscr{I} \circ \mathscr{P} \circ \mathscr{S}$, $\mathscr{F}_2$ is compact. Let $\mathscr{G}: E_1 \to H^1([0, l] \times [0, T])$ be defined by $\mathscr{G}(\phi^0, \phi^1) = y$. By conservation of energy and Poincaré's inequality, $\mathscr{G}$ is continuous. Let $\mathscr{J}$ be the embedding from $H^1([0, l] \times [0, T])$ into $L^2([0, l] \times [0, T])$. Then $\mathscr{J}$ is compact. Because $\mathscr{F}_3 = \mathscr{J} \circ \mathscr{G}$, $\mathscr{F}_3$ is compact. Hence $\mathscr{F}$ is indeed compact. Inequalities (33) and (34) imply, respectively, the following inequalities:

$$M_1 \| \mathscr{H}_2(\phi, \psi) \|^2 + \| \mathscr{F}(\phi, \psi) \|^2 \geqq \frac{K}{2} (T - T_1^*) \| (\phi, \psi) \|^2 \tag{35}$$

and

$$M_2 \| \mathscr{H}_3(\phi, \psi) \|^2 + \| \mathscr{F}(\phi, \psi) \|^2 \geqq \frac{K}{2} (T - T_1^*) \| (\phi, \psi) \|^2, \tag{36}$$

where

$$M_1 = \max \{ |\rho(x)| : x_1 \leqq x \leqq x_2 \} > 0,$$

$$M_2 = \max \{ |\sigma(x)| : x_1 \leqq x \leqq x_2 \} > 0$$

and we have used the equivalent norm

$$\| (\phi, \psi) \| = \left(\frac{1}{2} \int_0^l |\phi'|^2 + |\psi|^2 \, dx \right)^{1/2}$$

for the space $H_0^1[0, l] \times L^2[0, l]$. We then invoke the following result of functional analysis. We include its proof for completeness.

THEOREM 7. *Let X, Y, and Z be Hilbert spaces, and let $A: X \to Y$, $K: X \to Z$ be bounded linear operators. Suppose A is injective, K is compact, and there exists a constant $m > 0$ such that*

$$\| Ax \|^2 + \| Kx \|^2 \geqq m \| x \|^2 \quad \textit{for all } x \in X. \tag{37}$$

Then A is boundedly invertible. In other words, there exists a constant $m_1 > 0$ such that

$$\| Ax \|^2 \geqq m_1 \| x \|^2$$

for all $x \in X$.

Proof. If A is not boundedly invertible, then we can find a sequence $\{x_n\}$, $\| x_n \| = 1$ for all n, such that

$$\lim_{n \to \infty} \| Ax_n \| = 0.$$

Without loss of generality, we may assume that x_n converges weakly to some $\bar{x} \in X$. Now $\| Ax \|^2$, being a convex function in x, is weakly lower semicontinuous. Hence

$$\| A\bar{x} \|^2 \leqq \liminf_n \| Ax_n \|^2 = 0.$$

Because A is injective, this implies $\bar{x} = 0$. Since K is compact, we then have

$$\lim_{n \to \infty} Kx_n = K\bar{x} = 0.$$

Putting $x = x_n$ in (37), we have

$$\|Ax_n\|^2 + \|Kx_n\|^2 \geqq m\|x_n\|^2 = m \quad \text{for all } n.$$

Letting $n \to \infty$, we conclude that $0 \geqq m$, which is a contradiction. Hence A is boundedly invertible. □

Now we let $T_2^* = T_1^*$. By Proposition 3, $\mathscr{H}_2$ is injective if $T > T_2^*$. Since (35) holds, by Theorem 7 we conclude that $\mathscr{H}_2$ is boundedly invertible. This proves (5). Finally, we let T_3^* be such that the conclusion of Proposition 4 holds so that $\mathscr{H}_3$ would be injective. Note that we actually have $T_3^* > T_1^*$. Hence for $T > T_3^*$, (36) holds and hence $\mathscr{H}_3$ is, by Theorem 7, boundedly invertible. This completes the proof of inequalities (5) and (6).

5. From observability to controllability. In this section, we will prove the controllability results from the observability results. The duality relation between controllability and observability for distributed parameter systems was first noted by Dolecki and Russell [1]. Lions [4] developed it into the Hilbert Uniqueness Method (HUM). As we will see, the bounded invertibility of $\mathscr{H}_2$ and $\mathscr{H}_3$ gives us, respectively, the two exact controllability results in Theorem 1 with different regularity assumptions on the control function u.

The method of our proof is new. It is based on a relationship between a control problem and a corresponding minimization problem associated with its adjoint system.

Let us consider the following control system.

PROBLEM (I).

$$\rho \frac{\partial^2 z}{\partial t^2} = \frac{\partial}{\partial x}\left(\sigma \frac{\partial z}{\partial x}\right) + u, \qquad t > 0, \quad 0 \leqq x \leqq l,$$

$$z(x, 0) = \frac{\partial z(x, 0)}{\partial t} = 0, \qquad 0 \leqq x \leqq l, \tag{38}$$

$$z(0, t) = z(l, t) = 0, \qquad t > 0.$$

Given $(\phi^0, \phi^1) \in H_0^1[0, l] \times L^2[0, l]$, find $u \in L^2([x_1, x_2] \times [0, T])$ such that $z(x, T) = \phi^0(x)$, $(\partial z(x, T)/\partial t) = \phi^1(x)$, $0 \leqq x \leqq l$.

The following minimization problem is related to the above problem.

PROBLEM (I′). Minimize

$$J_1(\psi^0, \psi^1) = \frac{1}{2}\int_0^T \int_{x_1}^{x_2} \rho\left(\frac{\partial y}{\partial t}\right)^2 dx\, dt - \int_0^l \rho\psi^1\phi^1 + \sigma \frac{d\psi^0}{dx}\frac{d\phi^1}{dx}\, dx,$$

where

$$\rho \frac{\partial^2 y}{\partial t^2} = \frac{\partial}{\partial x}\left(\sigma \frac{\partial y}{\partial x}\right), \qquad 0 \leqq x \leqq l, \quad t > 0,$$

$$y(x, T) = \psi_0(x), \quad \frac{\partial y(x, T)}{\partial t} = \psi_1(x), \quad 0 \leqq x \leqq l, \tag{39}$$

$$y(0, t) = y(\ell, t) = 0, \qquad 0 \leqq t \leqq T.$$

$(\psi^0, \psi^1) \in H_0^1[0, l] \times L^2[0, l] = E_1$.

THEOREM 8. *If problem* (I′) *has an optimal solution* (ψ^0, ψ^1), *then* $u = \rho\partial y/\partial t$ *on* $[x_1, x_2] \times [0, T]$ *solves problem* (I).

Proof. If (ψ^0, ψ^1) is an optimal solution of problem (I'), then for any pair $(\eta^0, \eta^1) \in E_1$, we have

$$
\begin{aligned}
0 &= \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(J_1(\psi^0 + \varepsilon\eta^0, \psi^1 + \varepsilon\eta^1) - J_1(\psi^0, \psi^1)\right) \\
&= \int_0^T \int_{x_1}^{x_2} \rho \frac{\partial w}{\partial t} \frac{\partial y}{\partial t}\, dx\, dt - \int_0^l \rho\eta^1\phi^1 + \sigma \frac{d\eta^0}{dx} \frac{d\phi^0}{dx}\, dx
\end{aligned} \tag{40}
$$

where w satisfies

$$
\begin{aligned}
&\rho \frac{\partial^2 w}{\partial t^2} = \frac{\partial}{\partial x}\left(\sigma \frac{\partial w}{\partial x}\right), && 0 \leqq x \leqq l, \qquad t > 0, \\
&w(x, T) = \eta^0(x), \quad \frac{\partial w(x, T)}{\partial t} = \eta^1(x), && 0 \leqq x \leqq l, \\
&w(0, t) = w(\ell, t) = 0, && 0 \leqq t \leqq T.
\end{aligned} \tag{41}
$$

On the other hand,

$$
\begin{aligned}
&\frac{\partial}{\partial t} \int_0^l \rho \frac{\partial z}{\partial t} \frac{\partial w}{\partial t} + \sigma \frac{\partial z}{\partial x} \frac{\partial w}{\partial x}\, dx \\
&\quad = \int_0^l \rho \left(\frac{\partial^2 z}{\partial t^2} \frac{\partial w}{\partial t} + \frac{\partial z}{\partial t} \frac{\partial^2 w}{\partial t^2}\right) + \sigma \left(\frac{\partial^2 z}{\partial t \partial x} \frac{\partial w}{\partial x} + \frac{\partial z}{\partial x} \frac{\partial^2 w}{\partial t \partial x}\right) dx \\
&\quad = \int_0^l \frac{\partial}{\partial x}\left(\sigma \frac{\partial z}{\partial x}\right) \frac{\partial w}{\partial t} + u \frac{\partial w}{\partial t} + \frac{\partial z}{\partial t} \frac{\partial}{\partial x}\left(\sigma \frac{\partial w}{\partial x}\right) \\
&\qquad + \sigma \left(\frac{\partial^2 z}{\partial t \partial x} \frac{\partial w}{\partial x} + \frac{\partial z}{\partial x} \frac{\partial^2 w}{\partial t \partial x}\right) dx \\
&\quad = \int_0^l \frac{\partial}{\partial x}\left(\sigma \frac{\partial z}{\partial x} \frac{\partial w}{\partial t} + \sigma \frac{\partial z}{\partial t} \frac{\partial w}{\partial x}\right) + u \frac{\partial w}{\partial t}\, dx \\
&\quad = \int_0^l u \frac{\partial w}{\partial t}\, dx \\
&\quad = \int_{x_1}^{x_2} \rho \frac{\partial y}{\partial t} \frac{\partial w}{\partial t}\, dx.
\end{aligned}
$$

Integrating from $t = 0$ to $t = T$, we have

$$
\begin{aligned}
&\int_0^l \rho(x) \frac{\partial z(x, T)}{\partial t} \eta^1(x) + \sigma(x) \frac{\partial z(x, T)}{\partial x} \frac{d\eta^0(x)}{dx}\, dx \\
&\qquad = \int_0^T \int_{x_1}^{x_2} \rho \frac{\partial y}{\partial t} \frac{\partial w}{\partial t}\, dx.
\end{aligned} \tag{42}
$$

From (40) and (42), we have

$$
\begin{aligned}
&\int_0^l \rho(x) \frac{\partial z(x, T)}{\partial t} \eta^1(x) + \sigma(x) \frac{\partial z(x, T)}{\partial x} \frac{d\eta^0(x)}{dx}\, dx \\
&\qquad = \int_0^T \int_{x_1}^{x_2} \rho \frac{\partial w}{\partial t} \frac{\partial y}{\partial t}\, dx\, dt = \int_0^l \rho\eta^1\phi^1 + \sigma \frac{d\eta^0}{dx} \frac{d\phi^0}{dx}\, dx.
\end{aligned} \tag{43}
$$

Because (η^0, η^1) can be arbitrary, it follows that

$$z(x, T) = \phi^0(x), \qquad \frac{\partial z(x, T)}{\partial t} = \phi^1(x). \qquad \square$$

Note that a *sufficient* condition for problem (I′) to have a solution for any pair (ϕ^0, ϕ^1) is that the quadratic part of J_1 satisfies

$$\frac{1}{2}\int_0^T \int_{x_1}^{x_2} \rho \left(\frac{\partial y}{\partial t}\right)^2 dt \geqq K\|(\psi^0, \psi^1)\|^2 \tag{44}$$

because in that case, J_1 would be convex and coercive for any (ψ^0, ψ^1). But inequality (44) follows from Theorem 2 (for $T > T_2^*$). This completes the proof of the first part of Theorem 1 ($i = 1$ case).

The proof of the second part ($i = 2$ case) is similar. We consider the following two problems.

PROBLEM (II). Given $(\phi^0, \phi^1) \in E_2 = L^2[0, l] \times H^{-1}[0, l]$, find $u \in U_2 = L^2(0, T; H^{-1}[0, l])$; u vanishes outside $[x_1, x_2] \times [0, T]$ such that the solution z of (38) satisfies

$$z(x, T) = \phi^0(x), \quad \frac{\partial z(x, T)}{\partial t} = \phi^1(x) \quad \text{for } x \in [0, l].$$

PROBLEM (II′). Minimize

$$J_2(\psi_0, \psi_1) = \frac{1}{2}\int_0^T \int_{x_1}^{x_2} \sigma \left(\frac{\partial y}{\partial x}\right)^2 dx\, dt + \int_0^l \rho\psi^1\phi^0\, dx - \langle \phi^1, \rho\psi^0 \rangle_{H^{-1}[0,l] \times H_0^1[0,l]}$$

where y satisfies (39) and $(\psi^0, \psi^1) \in H_0^1[0, l] \times L^2[0, l]$.

THEOREM 9. *If problem* (II′) *has an optimal solution* (ψ^0, ψ^1), *then defining* $u \in L^2(0, T; H^{-1}[0, l])$ *by*

$$\int_0^T \langle u, v \rangle\, dt = \int_0^T \int_{x_1}^{x_2} \sigma \frac{\partial y}{\partial x} \frac{\partial}{\partial x}\left(\frac{v}{\rho}\right) dx\, dt \tag{45}$$

for all $v \in L^2(0, T; H_0^1[0, l])$, u *solves problem* (II).

Proof. If (ψ^0, ψ^1) is an optimal solution of problem (II′), then similar to the proof of Theorem 8, we have

$$0 = \int_0^T \int_{x_1}^{x_2} \sigma \frac{\partial y}{\partial x} \frac{\partial w}{\partial x}\, dx\, dt + \int_0^l \rho\eta^1\phi^0\, dx - \langle \phi^1, \rho\eta^0 \rangle_{H^{-1}[0,l] \times H_0^1[0,l]} \tag{46}$$

where w satisfies (41). Consider the function

$$F(t) = \left\langle \frac{\partial z}{\partial t}, \rho w \right\rangle_{H^{-1}[0,l] \times H_0^1[0,l]} - \int_0^l \rho z \frac{\partial w}{\partial t}\, dx$$

where z satisfies (38) with u defined by (45). Then

$$\begin{aligned} F'(t) &= \left\langle \frac{\partial^2 z}{\partial t^2}, \rho w \right\rangle - \int_0^l \rho z \frac{\partial^2 w}{\partial t^2}\, dx \\ &= \left\langle \frac{\partial}{\partial x}\left(\sigma \frac{\partial z}{\partial x}\right) + u, \rho w \right\rangle - \int_0^l z \frac{\partial}{\partial x}\left(\sigma \frac{\partial w}{\partial x}\right) dx \\ &= \langle u, \rho w \rangle. \end{aligned}$$

(We may first assume that η^0 and η^1 are sufficiently smooth and hence w is also sufficiently smooth so that all the expressions appearing in the above equality make sense. Then by approximating (η^0, η^1) by smooth functions and taking the limit, we have $F'(t) = \langle u, \rho w \rangle$ for any $(\eta^0, \eta^1) \in E_1$.) Integrating from $t = 0$ to $t = T$ and noting that $F(0) = 0$, we have

$$
\begin{aligned}
F(T) &= \int_0^T \langle u, \rho w \rangle \, dx \\
&= \int_0^T \int_{x_1}^{x_2} \sigma \frac{\partial y}{\partial x} \frac{\partial w}{\partial x} \, dx \, dt \quad \text{(by (45))}.
\end{aligned} \tag{47}
$$

Combining (46) and (47), we have

$$
F(T) = \left\langle \left. \frac{\partial z}{\partial t} \right|_{t=T}, \rho \eta^0 \right\rangle - \int_0^l \rho z|_{t=T} \eta^1 \, dx = \langle \phi^1, \rho \eta^0 \rangle - \int_0^l \rho \eta^1 \phi^0 \, dx.
$$

Since η^0, η^1 can be arbitrary, this proves that

$$
z|_{t=T} = \phi^0 \quad \text{and} \quad \left. \frac{\partial z}{\partial t} \right|_{t=T} = \phi^1. \qquad \square
$$

Again, Theorem 2 implies that J_2 is coercive for any $(\phi_0, \phi^1) \in E_2$ when $T > T_3^*$. So problem (II′) always has an optimal solution, hence by Theorem 9, problem (II) always has a solution. This proves the second part of Theorem 1 ($i = 2$ case).

6. Concluding remarks. We have proved the exact controllability of the one-dimensional wave equation with locally distributed control. We have introduced a method for choosing the "right multiplier" for a problem as a solution of an ordinary differential equation with parameters that depend on the parameters of the system. Second, we have given a proof for the part in which observability implies controllability through the introduction of a minimization problem related to the control problem. For many control problems, we can find similarly related minimization problems. For readers familiar with the examples of the systems in [4], it should not be a difficult exercise to write the corresponding minimization for each such problem.

REFERENCES

[1] S. DOLECKI AND D. L. RUSSELL, *A general theory of observations and control*, SIAM J. Control Optim., 15 (1977), pp. 185-220.

[2] L. F. HO, *Observabilité frontière de l'équation des ondes*, C.R. Acad. Sci. Paris, 302 (1986), pp. 443-446.

[3] I. LASIECKA, J. L. LIONS, AND R. TRIGGIANI, *Non-homogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., 65 (1986), pp. 149-192.

[4] J. L. LIONS, *Exact controllability, stabilization and perturbation of distributed systems*, SIAM Rev., 30 (1988), pp. 1-68.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 749-788, July 1990

Invited Expository Article

This paper is another in the continuing series of expository papers that were invited by the editors. These papers undergo the same refereeing procedure as do research papers submitted directly by the authors, although the refereeing guidelines are modified to suit the largely expository nature of the paper. Due to the rapid recent technical development of a number of areas in control and optimization, many of the seminal papers are quite specialized and are readily accessible to a limited group of experts only. Moreover, the original motivations and practical importance of the ideas are sometimes difficult to find in the mathematical development. The purpose of these papers is to bring the ideas, techniques, and applications of a few selected areas to the attention of a wider audience, so that their basic importance can be more easily and widely appreciated.

A SURVEY OF VIABILITY THEORY*

JEAN-PIERRE AUBIN†

Abstract. Some theorems of viability theory which are relevant to nonlinear control problems with state constraints and state-dependent control constraints are motivated and surveyed. They all deal with viable solutions to nonlinear control problems, i.e., solutions satisfying at each instant given state constraints of a general and diverse nature.

Some classical results on controlled invariance of smooth nonlinear systems are adopted to the nonsmooth case, including inequality constraints bearing on the state and state-dependent constraints on the controls.

For instance, existence of a viability kernel of a closed set (corresponding to the largest controlled invariant manifold) is provided under general conditions, even when the zero-dynamics algorithm does not converge.

The concepts of slow and heavy viable solutions are introduced, providing concrete ways of regulating viable solutions, by closed-loop feedbacks and closed-loop dynamical feedbacks.

Viability theorems also allow the extension of Lyapunov's second method to nonsmooth observation functions and the construction of "best" Lyapunov functions. As an application, "fuzzy differential inclusion" is presented.

Proofs and complements can be found in [*Viability Theory*, to appear, 1991]. They rely on properties of differential inclusion (see [*Differential Inclusions*, Springer-Verlag, Berlin, New York, 1984]) and set-valued analysis, (see [*Set-Valued Analysis*, Birkhäuser, Basel, 1990]).

Key words. viability, invariance, controlled invariance, set-valued maps, regulation map, differential inclusion, fuzzy differential inclusion, Lyapunov stability, asymptotic stability, tracking, contingent cone, contingent derivative of a set-valued map, epicontingent derivative of a function

AMS(MOS) subject classifications. 26A27, 26A51, 26E25, 28B20, 28D05, 34A60, 34D, 39A, B, 49A52, 54C60, 54C65, 58C06, 58C07, 58C30, 93C15, 93C30

Introduction. Viability theory offers *mathematical metaphors of evolution of macrosystems* arising in biology, economics, cognitive sciences, games, and similar areas.

The mathematical machinery built during the last decade to study the evolution of such systems appears to be as relevant and useful for solving some problems[1] arising in nonlinear systems theory as differential geometry.[2] This is the reason for this survey to appear in this journal.

* Received by the editors April 10, 1989; accepted for publication (in revised form) September 26, 1989.

† Centre de Recherches de Mathématiques de la Décision, Universite de Paris-Dauphine, F-75775 Paris cx (16), France and International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria.

[1] Existence of *viability kernels* (and thus, "zero-dynamics") in the general case, explicit construction of feedbacks and dynamical feedbacks, best Lyapunov functions, asymptotic observability, tracking problems, decentralization, decomposition properties, are problems that can be solved by the tools provided by viability theory with at least the same success than those of differential geometry.

[2] Tangent spaces to manifolds being replaced by contingent cones to any subsets, vector fields by differential inclusions on a viability domain, local existence by global existence, zero dynamics by viability kernels, Lie brackets by Frankowska's second order variations (see [62], although set-valued analysis techniques applied to controllability, observability and optimal control are not surveyed here). Viability theory and set-valued analysis offer new tools to control scientists.

Before presenting the main theorems constituting viability theory in the framework of control systems (which was no part of the motivations of viability theory until recently), we briefly explain the origins and purposes of viability theory.

Two main, common features of such macrosystems hold our attention:

— *their lack of determinism*, i.e., the possibility at each moment of several evolutions which depend upon the state, or even the history of the evolution of the state of the system up to this moment (*les jeux ne sont jamais faits*).

This lack of determinism covers many different aspects: it may be due to "uncertainty"[3] to "disturbances" and "perturbations" of various kinds, or to errors in modeling due to the impossibility of a comprehensive description of the dynamics of the system.

In several instances, the dynamics of the system are related to certain "controls," which, in turn, are restricted by state-dependent constraints (closed loop systems).

— *the presence of viability constraints* that the state of the system must obey at each time.

In a nutshell, *the main purpose of viability theory is to explain possible viable evolutions of a system, determined by given nondeterministic dynamics and state constraints, to reveal the concealed feedbacks which allow the system to be regulated and provide selection mechanisms for implementing them.*

Contrary to *optimal control theory*, viability theory does not require a single decision-maker (or actor, or player) to "guide" the system by optimizing an *intertemporal* optimality criterion.[4] Furthermore, the choice (even conditional) of the controls is not made *once and for all* at some initial time, but *they can be changed at each moment to take into account possible modifications of the environment of the system*, allowing therefore for *adaptation* to viability constraints.

Finally, by not appealing to intertemporal criterion, *viability theory does not require any knowledge of the future*[5] (even of a stochastic nature). This is of particular importance when experimentation[6] is not possible or when the phenomenon under study is not periodic. For example, in biological evolution as well as in economics and other such macrosystems, *the dynamics of the system disappear and cannot be recreated.* Hence, forecasting or prediction of the future are not the issues addressed by viability theory.

However, the conclusions of the viability theorems allow us to reduce the choice of possible evolutions, or to single out impossible future events, or to provide explanation of some behaviors which do not fit any reasonable optimality criterion.

Indeed, an interesting consequence of this lack of dependency on the future is that viability theory replaces the familiar paradigm of selection procedures of available evolutions via *intertemporal optimization* criteria[7] that depend on the future. It does so

---

[3] No a priori knowledge of an underlying probability law on the state of events is made. Fuzzy viability provides models where the available velocities can be ranked through a membership cost function to take into account that some velocities are more likely to be chosen than others.

[4] The choice of which is open to question even in static models, even when multicriteria or several decision makers are involved in the model.

[5] Many macrosystems do involve myopic behavior; while they cannot take into account the future, they are constrained by the past.

[6] Experimentation, by assuming that the evolution of the state of the system starting from a given initial state for a same period of time will be the same whatever the initial time, allows one to translate the time interval back and forth, and, thus, to "know" the future evolution of the system.

[7] Which can be traced back to Sumerian mythology which is at the origin of Genesis: one Decision-Maker, deciding what is good and bad and choosing the best (fortunately, on an intertemporal basis, thus wisely postponing to eternity the verification of optimality), knowing the future, and having taken the optimal decisions, well, during one week . . .

by using selection procedures of *viable evolutions* obeying, at each moment, state constraints which depend upon the *present or the past.* (This does not exclude *anticipations*, which are extrapolations of past evolutions, constraining in the last analysis the evolution of the system to be a function of its history).

Nonetheless, selection through viability constraints may not be discriminating enough. Starting from any state at any instant, several viable solutions may be implemented by the system, including equilibria, which are stationary evolutions.[8]

Thus further selection mechanisms need to be devised and/or discovered. We advocate here a third feature to which a selection procedure must comply:

— *Inertia Principle*: which states that "*the controls are kept constant as long as viability of the system is not at stake.*"

Indeed, as long as the state of the system lies in the interior of the constraint set (the set of states satisfying viability constraints), any regularity control will work. Therefore, the system can maintain the control inherited from the past. This happens if the system obeys the inertia principle. Since the state of the system may evolve while the control remains constant, it may reach the viability boundary with an "outward" velocity. This event corresponds to a period of *crisis*: To resolve the crisis, the system must find another regulatory control such that the new associated velocity forces the solution back inside the constraint set.

Naturally, there are several procedures for selecting a viable control when viability is at stake. For instance, the selection at each instant of the controls providing viable evolutions with *minimal velocity* is an example that obeys this inertia principle. They are called "*heavy*" *viable evolutions*[9] in the sense of heavy trends in economics.

Heavy viable evolutions can be viewed as providing mathematical metaphors for the concept of *punctuated equilibrium* introduced in paleontology by Elredge and Gould.

On the mathematical side, viability theory contributed to vigorous renewed interest in the field of "differential inclusions," as well as an engine for the development of a differential calculus of set-valued maps.[10] Indeed, as it often occurs in mathematics, these techniques can be relevant to control theory of nonlinear systems: the viability property has been studied independently in control theory under the name of *controlled invariance* in the framework of the geometric approach linear and smooth nonlinear systems, and the concept of *viability kernel* is closely related to the concept of *zero-dynamics* studied by Byrnes and Isidori [27]–[30] and Krener [77].

The mathematical tools designed to answer the above questions can replace the standard geometrical tools and bypass many regularity requirements required on the constraint sets, which need only to be closed (or on the Lyapunov functions, which can be taken only lower semicontinuous).

We proceed in this introduction with a description of what we think are the most convincing results.

---

[8] It may be observed that the state of the system becomes increasingly robust the further it is from the boundary of the constraint set. Therefore, after some time has elapsed, only the parts of the trajectories furthest away from the viability boundary will remain. This fact may explain the apparent discontinuities ("missing links") and hierarchical organization arising from evolution in certain systems.

[9] When the controls are the velocities, heavy solutions are the ones with minimal acceleration, i.e., maximal inertia.

[10] One can say that by now the main results of functional analysis have their counterpart in what can be called *Set-Valued Analysis.* Only the results needed in this book will be presented. An exposition of Set-Valued Analysis can be found in the monograph [8] by J.-P. Aubin and H. Frankowska.

Consider the evolution of a control system with (multivalued) feedbacks:

$$\begin{cases} \text{i)} & x'(t) = f(x(t), u(t)) \\ \text{ii)} & u(t) \in U(x(t)) \end{cases}$$

where the state $x(\cdot)$ ranges over a finite-dimensional vector-space X and the control $u(\cdot)$ over the finite-dimensional vector-space Z. The set-valued map $U: X \rightsquigarrow Z$ may be called an "a priori feedback." It describes the *dependence* of admissible controls on the actual state of the system. Such dependence arises quite often in many problems, and appears as soon as state constraints have to be satisfied, as the Viability Theorem will show later.

A solution to this system is a function $t \to x(t)$ satisfying this system for some control $t \to u(t)$.

State constraints (here also called viability constraints) are described in the last analysis by a closed subset K of the state space: The state of the system must remain in K; outside of K, the state of the system is no longer viable.

A subset K enjoys the *viability property* (for the control system described by f and U) if for every initial state $x_0 \in K$, there exists at least one solution to the system starting at x_0 which is *viable* in the sense that

$$\forall t \in [0, T], \quad x(t) \in K.$$

For linear control systems, this property has been introduced under the name of "controlled invariance" in [11], [90], [120]. See also [117]-[119] for instance. This property has then been extended to nonlinear systems in [26]-[30], [70], [75], [74], [116].

The first task is to characterize the subsets having this property, without solving the system and checking the existence of viable solutions for each initial state.

We cannot be content with constraint sets that are smooth manifolds, because inequality constraints would thereby be ruled out. We shall choose from among the many ways of "implementing" the concept of "tangency" for any subset K the one suggested by Bouligand fifty years ago: a direction v is *contingent to K at $x \in K$* if it is a limit of a sequence of directions v_n such that $x + h_n v_n$ belongs to K for some sequence $h_n \to 0+$. The collection of such directions, which are in some sense "inward," constitutes a closed cone $T_K(x)$, called the *contingent cone*[11] to K at x.

We then associate with the dynamical system (described by f and U) and with the state constraints (described by K) the (*set-valued*) *regulation map* R_K. It maps any state x to the subset $R_K(x)$ consisting of controls $u \in U(x)$ which are *viable* in the sense that $u \in U(x)$ and $f(x, u)$ is contingent to K at x.

If, for every $z \in K$, there exists at least one viable control $u \in R_K(x)$, we then say that K is a *viability domain* of the control system with dynamics described by both f and U.

The Viability Theorem we mentioned earlier holds true for a rather large class of systems: beyond some weak technical conditions, the only severe restriction is that, for each state x, the set of velocities $f(x, u)$ when u ranges over $U(x)$ is *convex*. This includes control systems which are affine with respect to the control. From now on, we assume that the systems under investigation satisfy these assumptions.

The basic viability theorem states that for such systems, *a closed subset K enjoys the viability property if and only if K is a viability domain.*

---

[11] Replacing the linear structure underlying the use of tangent spaces by the contingent cone is at the root of *Set-Valued Analysis.*

Interesting subsets such as equilibrium points, trajectories of periodic solutions and the limit sets of solutions are examples of closed viability domains. Actually, equilibrium points $\bar{x}$, which are solutions to

$$f(\bar{x}, \bar{u}) = 0 \quad \text{where } \bar{u} \in U(\bar{x})$$

are the smallest viability domains, the ones reduced to a single point, since, being *stationary states*, their velocities $f(\bar{x}, \bar{u})$ are equal to zero.

There exists a basic and curious link between viability theory and general equilibrium theory: *every compact convex viability domain contains an equilibrium point.* This statement is a version of the *Brouwer Fixed Point Theorem*, the cornerstone of nonlinear analysis, which finds here a particularly relevant formulation (viability implies stationarity).

When a closed subset K is not a viability domain, we can state that *there exists a largest closed viability domain contained in K.* This domain will be denoted Viab (K) and called the *viability kernel*[12] of K. It may be empty.

Contrary to the case of smooth systems and linear constraints, the existence of the viability kernel (largest controlled invariant manifold) is not obtained through the zero-dynamics algorithm, a generalization by Byrnes and Isidori [27]–[30] of the Basile–Marro and Silverman algorithms devised in the linear case. We shall provide a simple counterexample which shows that this algorithm does not converge to a viability kernel when inequality constraints are involved.

The Viability Theorem also provides a *regulation law* for regulating the system in order to maintain the viability of a solution: When K is a viability domain, the viable solutions $x(t)$ are regulated by viable "open loop controls" $u(t)$ through the regulation law:

$$\text{for almost all } t, \quad u(t) \in R_K(x(t)).$$

The multivaluedness of the regulation map is an indicator of the "robustness" of the system: The larger the set $R_K(x(t))$, the larger the set of disturbances which do not destroy the viability of the system!

Observe that solutions to a control system are solutions to the differential inclusion $x'(t) \in F(x(t))$ where, for each state x, $F(x) := f(x, U(x))$ is the subset of feasible velocities. Conversely, a differential inclusion is an example of a control system in which the controls are the velocities ($f(x, u) = u$ and $U(x) = F(x)$).

Observe also that whenever feasible controls obey state-dependent constraints, it can no longer be regarded as a family of differential equations parametrized by an open loop control $u(\cdot)$, but as a differential inclusion.

As far as servomechanisms are concerned, the question arises of how to build mechanisms for selecting a *closed loop* control $\hat{u}(x)$ in $R_K(x)$ for each state x. Such a single-valued map $\hat{u}(\cdot)$ allows the system to automatically associate with any state $x(t)$ the control $\hat{u}(x(t))$ which produces a viable solution through the differential equation

$$x'(t) = f(x(t), \hat{u}(x(t)))$$

An interesting example of closed loop control is provided by *slow solutions.* These are the solutions regulated by the controls $u^0(x) \in R_K(x)$ with minimal norm. Despite the fact that $u^0(\cdot)$ is not necessarily continuous, we shall prove that the above differential

[12] This concept of viability kernel happens to be a quite efficient mathematical tool.

equation still has solutions. For instance, when the controls are the velocities of the system, viable solutions with *velocities of minimal norm* are implemented by such a selection procedure. This is why they are called *slow solutions.*

Such selection procedures by closed loop controls answer many engineering control problems, but are not adequate for another type of system arising in economics and biology which motivated viability theory, where we are looking for selection procedures which obey the *inertia principle*: *Keep the controls constants as long as the viability is not at stake.*

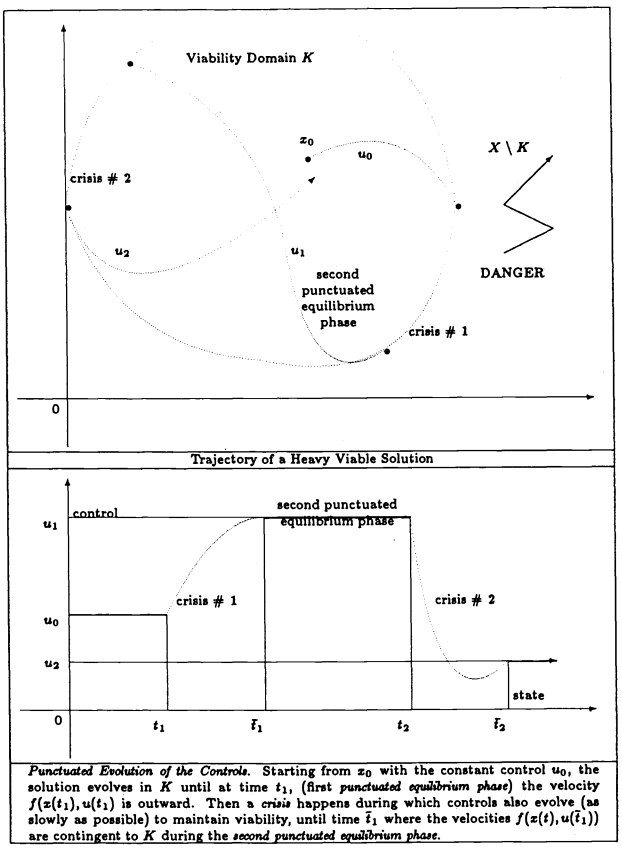
We can reformulate it by saying that if the derivative of a viable open loop control $u(\cdot)$ is equal to 0, this control is the one which is chosen and implemented.

This raises several questions. The first one concerns controls which are smooth (at least, differentiable almost everywhere). This issue may even be relevant for engineering problems, where the lack of continuity of controls $u(t) := \hat{u}(x(t))$ can be damaging.

The second one deals with the problem of differentiating the regulation law.

The third is to find selections (called *dynamical closed loops*) of the derivative of the regulation map, with which we obtain a system of differential equations which govern the *smooth* viable evolution of both the state and the control.

We see at once that this program requires a concept of derivative of a set-valued map and a chain rule formula in order to differentiate the regulation law.



Punctuated Evolution of the Controls. Starting from x_0 with the constant control u_0, the solution evolves in K until at time t_1, (first *punctuated equilibrium phase*) the velocity $f(x(t_1), u(t_1))$ is outward. Then a *crisis* happens during which controls also evolve (as slowly as possible) to maintain viability, until time $\bar{t}_1$ where the velocities $f(x(t), u(\bar{t}_1))$ are contingent to K during the *second punctuated equilibrium phase.*

FIG. 1. *Heavy viable solutions.*

The idea behind the construction of a differential calculus of set-valued maps is simple and goes back to the very origins of differential calculus, when Pierre de Fermat[13] introduced, in the first half of the seventeenth century, the concept of a tangent to the graph of a function: *the tangent space to the graph of a function f at a point (x, y) of its graph is the line of slope $f'(x)$, i.e., the graph of the linear function $u \mapsto f'(x)u$.*

Consider now a set-valued map $F: X \rightsquigarrow Y$, which is characterized by its graph (the subset of pairs (x, y) such that y belongs to $F(x)$): *The contingent cone to the graph of F at the point (x, y) of its graph is the graph of the contingent derivative of the set-valued map F at a point (x, y).* The contingent derivative is a set-valued map from X to Y denoted by $DF(x, y)$. Contingent derivatives keep enough properties of the derivatives of smooth functions to be quite efficient. They enjoy a rich calculus, and they enable such basic theorems of analysis as the inverse function theorem to be extended to the set-valued case.

The chain rule is an example of a property which is still true in this framework: Assume that we start from a "smooth state," producing a viable solution $x(t)$ and a viable control $u(t)$ which are both differentiable (almost everywhere). Then we can "differentiate" the regulation law and obtain a "first order regulation law":

$$\text{for almost all } t, \quad u'(t) \in DR_K(x(t), u(t))(x'(t)).$$

Heavy viable solutions are then the ones which are regulated by the controls whose velocities have minimal norm in the set

$$DR_K(x(t), u(t))(f(x(t), u(t))).$$

For instance, when the control is the velocity of the system, we choose in this way viable solutions with *acceleration of minimal norm*, i.e., accelerations with maximum inertia. This is why these solutions are called *heavy solutions.* This point of view leads to the introduction of *viability niches* $N(u)$ associated with controls u. These are (possibly empty) subsets consisting of states x such that the zero velocity belongs to $DR_K(x, u)(f(x, u))$. In such a viability niche $N(u)$, the state can evolve while being regulated by the stationary control u.

Although we shall present viability theory in the framework of control systems with state constraints and state-dependent control constraints for the readers of SIAM, we recall that these problems were also *motivated* in the first place by systems arising in "soft sciences" such as economics and biology.

In economics, when we can replace the fundamental Walrasian model[14] of resource allocations by a decentralized dynamical model in which the role of the controls is

[13] Fermat was one of the most important innovators in the history of mathematics. Newton himself recognized explicitly that he got the hint of the differential calculus from Fermat's method of building tangents devised half a century earlier. Fermat was also the one who discovered that the derivative of a (polynomial) function vanishes when it reaches an extremum. (This is Fermat's Rule, which remains the main strategy for obtaining necessary conditions of optimality, from mathematical programming to calculus of variations to optimal control.) Fermat also was the first to discover the "principle of least time" in optics, the prototype of the variational principles governing so many physical and mechanical laws. He shared independently with Descartes the invention of analytic geometry and with Pascal the creation of the mathematical theory of probability. He was on top of that a poet, a linguist, a lawyer and, if it has to be recalled, the author of the Fermat Theorem.

[14] Most static models of mathematical economics are based in the last analysis on *general equilibrium theory.* They can be reformulated in a dynamical framework by changing slightly the underlying dynamical system. (Walrasian "tâtonnement," which does not produce viable solutions, except when they reach an equilibrium.)

played by the prices[15] (as well as coalitions of consumers, interest rates, and so forth), and the regulation law can be interpreted as the behavior of Adam Smith's invisible hand choosing the prices as a function of the allocations. It is possible that among these viable prices, the market (or even a planning bureau) would have a tendency to choose heavy solutions.

In the case of cooperative games, coalitions of players may play the role of controls: each coalition acts on the environment by changing it through a dynamical system. Here, a coalition is described by the players's rate of participation, positive or negative, according to their cooperative or anticooperative behavior. The regulation law provides in this case an explanation of the evolution of coalitions and alliances.

In the noncooperative framework, viability constraints describe power relations among players, each player associating with each state a subset in which the other players are confined to choosing their own states. Strategies take the role of controls, and we often observe that the inertia principle is operative. The choice of viable strategies (or of their velocities) can be made, at each instant and in a myopic way, by standard game theoretical mechanisms, in such a way to comply with the inertia principle.

In sociology, a society can be interpreted in this framework as a set of individuals subject to viability constraints which maintain an organization needed for their survival. Laws and other cultural codes are then devised for providing each individual with psychological and economical means of survival as well as guidelines for avoiding conflicts. These cultural codes play the role of controls. The regulation law may represent the evolution of cultural codes for maintaining society's viability, the evolution of which obeys the inertia principle. This may account for the small number of them and the robustness of religions, ideologies, and scientiflc paradigms, and explain the phenomena of massive conversions to new cultural codes.

In cognitive sciences, the state describes the sensory-motor couple of the cognitive system, while the control translates into what could be called a conceptual control (which is the synaptic matrix in neural networks). The state and control are related by a pattern recognition mechanism which recognizes the (variations of) the perception of the action of the automaton on the environment. The regulation law provides a learning process that goes beyond simple stimulus-response processes: it associates with each sensory-motor state a subset of (learned) conceptual controls. It seems that in this case again, the inertia principle is at work.

Outline of the survey. We concentrate our survey on the basic viability theorems in the framework of differential inclusions (§ 1) and of the regulation of control systems in § 2.

We chose to give an example of applications of viability techniques to issues related to the asymptotic behavior of solutions to differential inclusions, such as solutions increasing along a preorder, comparison of solutions, tracking problems and asymptotic stability. We devote in particular a short section to differential inclusions the right-hand side of which are "fuzzy sets," assigning to each admissible velocity a membership cost.

We were forced by lack of space to leave aside other important issues such as time-dependent viability constraints (viability tubes) and their potential use in solving

[15] And other fiduciary goods for which the scarcity constraint can be transgressed. Unlike physical goods, they are limited only by measures dictated by the trust (or, rather, the tolerance) of the agents. Any disequilibrium that cannot exist in physical goods can then be transferred to the fiduciary goods.

the target problems, functional viability, where the viability constraints depend upon the history of the solution, extension of the viability theorem to parabolic partial differential inclusions and distributed systems, etc.

Viability can be used in differential games as we do in control problems in the second section of this survey, opening many interesting problems (see [9]).

The connections with problems arising in economics, biology, and cognitive sciences are naturally more metaphorical, but interesting enough to motivate many problems of viability theory.

1. Viability theorems for differential inclusions. In all this paper, X, Y, Z denote finite-dimensional vector-spaces, except an explicit mention to the contrary.

1.1. The viability property. Let us describe the (nondeterministic) dynamics of the system by a set-valued map F from the state space X to itself. We consider initial value problems (or Cauchy problems) associated to differential inclusion

$$\text{for almost all } t \in [0, T], \quad x'(t) \in F(x(t)) \tag{1}$$

satisfying the initial condition $x(0) = x_0$. We have first to agree on what we shall call a solution to such a differential inclusion.

In the case of differential equations, there is no ambiguity since the derivative $x'(\cdot)$ of a solution $x(\cdot)$ to a differential equation $x'(t) = f(t, x(t))$ inherits the properties of the map f and of the function $x(\cdot)$. This is no longer the case with differential inclusions. Hence, we shall look for solutions among *absolutely continuous functions.*

DEFINITION 1.1 (viability and invariance properties). Let K be a subset of Dom (F). A function $x(\cdot)$ from $[0, T]$ to X is called *viable* if for all $t \in [0, T]$, $x(t) \in K$. We shall say that K enjoys the local *viability property or controlled invariance* (for the set-valued map F) if for any initial state x_0 in K, there exist $T > 0$ and a *viable* solution on $[0, T]$ to differential inclusion (1) starting at x_0. It enjoys the global viability property (or, simply, the viability property) if we can take $T = \infty$.

The subset K is said to be *invariant or conditionally invariant under* F if for any initial state x_0 of K, all solutions to differential inclusion (1) (defined on the domain of F) are *viable* in K.

Remark. We should emphasize that the concept of viability depends only on the behavior of F on K whereas invariance *depends upon the behavior of* F *on the domain* Dom (F) *outside of* K.

1.2. Set-valued maps. We unfortunately need to recall some definitions about set-valued maps which may be known by many readers, who should then skip this subsection.

DEFINITION 1.2. If X and Y are metric spaces, a set-valued map F from X to Y is characterized by its *graph* Graph (F), subset of the product space $X \times Y$ defined by

$$\text{Graph}\,(F) := \{(x, y) \in X \times Y \mid y \in F(x)\}.$$

We shall say that $F(x)$ is the *image* or the *value* of F at x. A set-valued map is said to be *nontrivial* if its graph is not empty, i.e., if there exists at least an element $x \in X$ such that $F(x)$ is not empty.

We say that it is *strict* if all its images $F(x)$ are not empty. The *domain* of F is the subset Dom (F) of elements $x \in X$ such that $F(x)$ is not empty. The *image* Im (F)

of F is the union of the images (or values) $F(x)$ when x ranges over X. The *inverse* F^{-1} of F is the set-valued map from Y to X defined by

$$x \in F^{-1}(y) \quad \text{if and only if } y \in F(x).$$

If K is a subset of X, we denote by $F|_K$ its *restriction* to K, defined by

$$F|_K(x) := \begin{cases} F(x) & \text{if } x \in K \\ \varnothing & \text{if } x \notin K. \end{cases}$$

The *ball of radius r around K* is denoted by B_X or B when there is no ambiguity; we set $B_X(K, r) = K + rB_X$.

DEFINITION 1.3. A set-valued map $F: X \rightsquigarrow Y$ is called *upper semicontinuous* at $x \in \text{Dom}(F)$ if and only if

$$\forall \varepsilon > 0, \quad \exists \eta > 0 \,|\, \forall y \in B_X(x, n), \quad F(y) \subset B_Y(F(x), \varepsilon)$$

It is said to be *upper semicontinuous* on X if and only if it is upper semicontinuous at any point of Dom (F).

We shall say that a set-valued map $F: X \rightsquigarrow Y$ is *lower semicontinuous at* $x \in$ Dom (F) if and only if for all $y \in F(x)$ and for all sequence of elements x_n converging to x, there exists a sequence of elements $y_n \in F(x_n)$ converging to y. It is said to be *lower semicontinuous* on X if it is lower semicontinuous at every point $x \in$ Dom (F).

We shall say that a set-valued map is *continuous* at x if it is *both* upper semicontinuous and lower semicontinuous, and that it is *continuous* if and only if it is continuous at every point of Dom (F).

We shall say that F is *closed* if and only if its graph is closed.

Unfortunately, there exist set-valued maps which enjoy one property without satisfying the other. However, the graph of an upper semicontinuous set-valued map $F: X \rightsquigarrow Y$ with closed values is closed. The converse is true if we assume that Y *is compact.*

Example. Parametrized Set-Valued Maps. Let us consider three metric spaces X, Y, and Z, a set-valued map $U: X \rightsquigarrow Z$ and a single-valued map f: Graph $(U) \to Y$. We associate with these data the set-valued map $F: X \rightsquigarrow Y$ defined by

$$\forall x \in X, \quad F(x) := \{f(x, u)\}_{u \in U(x)}.$$

Let us assume that f is *continuous* from Graph (U) to Y.

— If U is lower semicontinuous, so is F.

— If U is upper semicontinuous with compact values, so is F.

1.3. Contingent cones. We provide the definition of the *contingent cone* with which we shall characterize the viability property. We denote for that purpose by $d_K(y)$ the distance of y to K, defined by $d_K(y) := \inf_{z \in K} \|y - z\|$.

DEFINITION 1.4. Let K be a nonempty subset of X and x belong to K. The *contingent cone* to K at x is the set

$$T_K(x) = \left\{ v \in X \;\middle|\; \liminf_{h \to 0+} \frac{d_K(x + hv)}{h} = 0 \right\}.$$

We shall say that a subset K of X is *sleek* at $x \in K$ if the set-valued map

$$K \ni x' \rightsquigarrow T_K(x') \text{ is lower semicontinuous at } x.$$

We shall say that it is *sleek* if and only if it is sleek at every point of K.

We see easily that for all $x \in \text{Int}(K)$, $T_K(x) = X$ (the converse is true when K is sleek at x) and that when K is a differential manifold, the contingent cone $T_K(x)$ coincides with the tangent space to K at x.

TABLE 1
Properties of contingent cones.

(1) ▷ If $K \subset L$, then $T_K(x) \subset T_L(x)$
(2) ▷ If $K_i \subset X$, $(i = 1, \cdots, n)$, then
$T_{\cup_{i=1}^n K_i}(x) = \cup_{i \in I(x)} T_{K_i}(x)$, where $I(x) := \{i \mid x \in \overline{K_i}\}$
(3) ▷ If $K_i \subset X_i$, $(i = 1, \cdots, n)$, then
$T_{\prod_{i=1}^n K_i}(x_1, \cdots, x_n) \subset \prod_{i=1}^n T_{K_i}(x_i)$
(4) ▷ If $g \in \mathscr{C}^1(X, Y)$, if $K \subset X$ and $M \subset Y$, then
$g'(x)(T_K(x)) \subset T_{g(K)}(x)$ and $T_{g^{-1}(M)}(x) \subset g'(x)^{-1} T_M(g(x))$
(5) ▷ If $L \subset X$ and $M \subset Y$ are *closed sleek* subsets and $f \in \mathscr{C}^1(X, Y)$ is a continuously differentiable map such that the *transversality condition* $f'(x)T_L(x) - T_M(x) = Y$ holds true, then
$T_{L \cap f^{-1}(M)}(x) = T_L(x) \cap f'(x)^{-1} T_M(f(x))$

THEOREM 1.5 (tangent cones of sleek subsets). *If a closed subset K is sleek at $x \in K$, then the contingent cone is convex.*

Any closed convex subset is sleek and its contingent cone $T_K(x)$ coincides with the tangent cone of convex analysis, which is the closed cone spanned by $K - x$.

We summarize in Table 1 the properties of the contingent cones to subsets.

DEFINITION 1.6 (viability domain). Let $F: X \rightsquigarrow X$ be a nontrivial[16] set-valued map. We shall say that a subset $K \subset \text{Dom}(F)$ is a *viability domain* of F if and only if

$$\forall x \in K, \quad F(x) \cap T_K(x) \neq \emptyset$$

and that it is an *invariance domain* if and only if

$$\forall x \in K, \quad F(x) \subset T_K(x).$$

Since the contingent cone to a singleton is obviously reduced to 0, we observe that a singleton $\{\bar{x}\}$ is a viability domain if and only if $\bar{x}$ is an equilibrium of F, i.e., a stationary solution to the inclusion $0 \in F(\bar{x})$. In other words, the equilibria of a set-valued map provide the first examples of viability domains, actually, the *minimal viability domains.*

Remark. If K is a viability domain of a set-valued map F, the subset

$$D := \bigcap_{x \in K} (T_K(x) - F(x))$$

is the subset of disturbances of the system which do not destroy the fact that K remains a viability domain, because K remains a viability domain of any set-valued map $x \rightsquigarrow F(x) + d(x)$ where $x \mapsto d(x)$ maps K into D.

1.4. Statements of the viability theorems. Viability theorems hold true for the class of nontrivial upper semicontinuous set-valued maps with nonempty compact convex images (see Definition 1.3). We observe that the only truly restrictive condition is the *convexity* of the images of these set-valued maps, since the continuity requirements are minimal:[17]

[16] See Definition 1.3 below.

[17] But we cannot dispense with it, as the following counter example shows. Let us consider $X := \mathbf{R}$, $K := [-1, +1]$ and the set-valued map $F: K \rightsquigarrow \mathbf{R}$ defined by $F(x) := -1$ if $x > 0$, $F(0) := \{-1, +1\}$ and $F(x) := +1$ if $x < 0$. Obviously, no solution to the differential inclusion $x'(t) \in F(x(t))$ can start from 0, since 0 is not an equilibrium of this set-valued map!

THEOREM 1.7 (Local Viability Theorem). *Let us consider a nontrivial upper semicontinuous set-valued map $F: X \rightsquigarrow X$ with compact convex images and a closed subset $K \subset \text{Dom } F$.*

Then K is a viability domain if and only if it enjoys the viability property. Actually, for any initial state $x_0 \in K$, there exist a positive T and a viable solution on $[0, T]$ to differential inclusion (1) such that either $T = \infty$ or $T < \infty$ and $\limsup_{t \to T-} \|x(t)\| = \infty$.

When $F \equiv f$ is single-valued, this theorem has been proved by Nagumo in 1942 and rediscovered fourteen times since.[18] The above set-valued version has been proved by Haddad in 1981 (see [64]).

Further adequate information—a priori estimates on the growth of F—allow us to exclude the case when $\limsup_{t \to T-} \|x(t)\| = \infty$. This is the case for instance when F is bounded on K, and, in particular, when K is bounded. More generally, we can take $T = \infty$ when F enjoys linear growth: for any $x \in K$, $\sup_{v \in F(x)} \|v\| \leqq c(\|x\| + 1)$.

We shall call *Peano maps* the nontrivial upper semicontinuous set-valued maps with nonempty compact convex images and with linear growth, or equivalently, the nontrivial closed set-valued maps with convex values and linear growth.

THEOREM 1.8 (Viability Theorem). *Let us consider a Peano map F from X to X and a closed subset $K \subset \text{Dom } F$. If K is a viability domain, then for any initial state $x_0 \in K$, there exists a viable solution on $[0, \infty]$ to differential inclusion (1).*

Let us now consider a sequence of closed viability domains of a set-valued map F and the following stability property: Is the upper limit[19] of these closed viability domains still a closed viability domain?

THEOREM 1.9 (stability of viability domains). *Let us consider a Peano map $F: X \rightsquigarrow X$. Then the upper limit of a sequence of closed viability domains of F is also a closed viability domain of F.*

Invariance property is characterized by invariant spaces in the case of Lipschitz maps (with nonconvex values).

THEOREM 1.10. *Let us assume that F is Lipschitz on the interior of its domain and has compact values. Then a closed $K \subset \text{Dom}(F)$ is invariant by F if and only if K is an invariance domain.*

Actually, the proof of the viability theorem yields local results. For that purpose, we need to introduce the "Dubovitsky-Miliutin tangent cone" $D_K(x)$ to K, which is defined by

$$v \in D_K(x) \quad \text{if and only if } \exists \varepsilon > 0, \quad \exists \alpha > 0 \quad \text{such that } x +]0, \alpha](v + \varepsilon B) \subset K$$

because the complement of the contingent cone $T_K(x)$ to K at $x \in \partial K$ is the "Dubovitsky-Miliutin cone" $D_{\hat{K}}(x)$ to the closure $\hat{K}$ of the complement of K.

PROPOSITION 1.11. *Let us consider a nontrivial upper semicontinuous set-valued map $F: X \rightsquigarrow X$ with compact convex images. Let $K \subset \text{Dom}(F)$ be closed with nonempty interior and $x_0 \in \partial K$. Then each of the following conditions implies the next one:*

(i) $F(x_0) \subset D_K(x_0)$

(ii) *for any solution starting from x_0, $\exists T > 0 \mid \forall t \in]0, T]$, $\quad x(t) \in \text{Int}(K)$*

---

[18] This does not imply that it is true, but stresses the lack of communication. "Everybody wants to teach, nobody wants to learn," Abel once complained bitterly.

[19] When K_n is a sequence of subsets of a metric space X, we say that

$$\limsup_{n \to \infty} K_n := \{y \in Y \mid \liminf_{n \to \infty} d(y, K_n) = 0\}$$

is its *upper limit*. It is the closed subset of cluster points of sequences of elements of K_n.

(iii) $\exists$ *a sequence* $x_n \in \partial K$ *converging to* x_0 *such that* $F(x_n) \subset D_K(x_n)$.

As a consequence, we obtain the following theorem.

THEOREM 1.12 (Strict Invariance Theorem). *Let us consider a nontrivial upper semicontinuous set-valued map* $F: X \rightsquigarrow X$ *with compact convex images and assume that the interior* K *is not empty. If*

$$\forall x \in \partial K, \quad F(x) \subset D_K(x)$$

then, for any initial state x_0 *in the boundary* ∂K *of* K, *for any solution to differential inclusion* (1) *starting from* x_0, *there exists* $T > 0$ *such that it remains in the interior of* K *on* $]0, T]$.

We denote by $\mathcal{S}(x_0)$ or by $\mathcal{S}_F(x_0)$ the (possibly empty) set of solutions to differential inclusion (1) and we call the set-valued map $\mathcal{S}$ defined by $\mathrm{Dom}(F) \ni x \mapsto \mathcal{S}(x)$ the *solution map* of F (or of differential inclusion (1)).

THEOREM 1.13 (continuity of the solution map). *Let us assume that* $F: X \rightsquigarrow X$ *is a Peano map. The solution map* $\mathcal{S}$ *is upper semicontinuous with compact images from its domain to the space* $\mathcal{C}(0, \infty; X)$ *of continuous functions* (*supplied with the compact convergence topology*).

DEFINITION 1.14 (viability and invariance kernel). Let K be a subset of the domain of a set-valued map $F: X \rightsquigarrow X$. We shall say that the largest closed viability domain contained in K (which may be empty) is the *viability kernel* of K and denote it by $\mathrm{Viab}_F(K)$ or, simply, $\mathrm{Viab}(K)$. The largest closed invariance domain contained in K, which we denote by $\mathrm{Inv}_F(K)$ or $\mathrm{Inv}(K)$, is called the *invariance kernel* of K (or, for smooth systems, the largest controlled invariant submanifold).

THEOREM 1.15. *Let us consider a Peano map* $F: X \rightsquigarrow X$. *Let* $K \subset \mathrm{Dom}(F)$ *be closed. Then the viability kernel of* K *exists* (*possibly empty*) *and is the subset of initial states such that at least one solution starting from them is viable in* K.

Let us assume that F *is Lipschitz on the interior of its domain and has compact values. For any closed subset* $K \subset \mathrm{Int}(\mathrm{Dom}(F))$, *there exists an invariance kernel* (*possibly empty*) *of* K. *It is the subset of initial states such that all solutions starting from them are viable in* K.

The viability kernels may inherit properties of both F and K. For instance, if the graph of F and the subset K are convex, so is the viability kernel of K. If F is a closed convex process (i.e., its graph is a closed convex cone) and if K is a closed convex cone, the viability kernel is a closed convex cone.

In general, viability kernels are not necessarily connected.

Remark. The zero dynamics algorithm has been devised to obtain the viability kernel of closed subsets defined by equality constraints, i.e., subsets of the form $K := h^{-1}(0)$ where h is a map from X to a finite dimensional vector-space Y. It is shown to converge for linear control systems (see [11], [107]) and for smooth nonlinear control systems (see [27]–[29]. In this framework, viability property is called *controlled invariance* and the restriction of the control system to the viability kernel is called *zero dynamics*).

In the general case, let us consider a closed subset K of the domain of a set-valued map $F: X \rightsquigarrow X$.

We start with $K_0 := K$ and we construct

$$K_1 = \mathrm{Dom}(R_{K_0}) \quad \text{where } R_{K_0}(x) := F(x) \cap T_K(x).$$

Since the viability kernel $\mathrm{Viab}_F(K)$ is contained in K and since $T_L(x) \subset T_K(x)$ whenever $K \subset L$, we infer that $\mathrm{Viab}_F(K) \subset K_1$.

Assume that a decreasing sequence of subsets K_i satisfying $\text{Viab}_F(K) \subset K_i \subset K_{i-1} \subset K$ has been defined up to n. We then set $R_{K_n}(x) := F(x) \cap T_{K_n}(x)$, define $K_{n+1} := \text{Dom}(R_{K_n})$ and we observe that $\text{Viab}_F(K) \subset K_{n+1}$.

Therefore

$$\text{Viab}_F(K) \subset \bigcap_{n=0}^{\infty} K_n.$$

The problem is to show that equality holds true. Several requirements have to be met to solve the problem. The first one is that the subsets K_n should be closed. The second is that the upper limit of the contingent cones $T_{K_n}(x)$ is contained in the contingent cone to the upper limit of the subsets K_n (which, in this case, is the intersection of the decreasing sequence of the subsets K_n).

These conditions are not met for finding the viability kernel of $K := [0, 1] \times \mathbf{R}$ for the system $F(x, v) := \{v\} \times cB$ since $K_0 = \{0\} \times \mathbf{R}_+ \cup]0, 1[\times \mathbf{R} \cup \{1\} \times \mathbf{R}_-$, $K_1 = K_0$ and since the viability kernel is shown in Fig. 2.

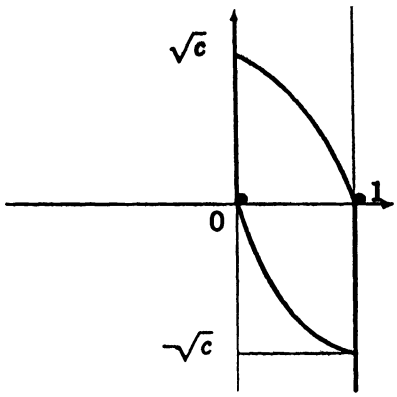


FIG. 2. *Viability kernel of* $[0, 1] \times \mathbf{R}$ *for* $F(x, v) := \{v\} \times cB$.

The viability kernel contains for instance all the limit sets $L(x)$ of the solutions $x(\cdot)$ to differential inclusion (1), defined by

$$L(x(\cdot)) := \bigcap_{T>0} cl(x([T, \infty[)).$$

THEOREM 1.16. *Let us consider a Peano map* $F : X \rightsquigarrow X$. *Then the limit sets of the solutions to differential inclusion* (1) *are* closed viability domains.

In particular, the limits of solutions to differential inclusion (1), *when they exist, are* equilibria *of F and the trajectories of periodic solutions to the differential inclusion* (1) *are also closed viability domains.*

Naturally, if a subsequence $x'(t_n)$ converges to 0, then a subsequence $x(t_{n_k})$ converges to an equilibrium of F. Actually, this statement can be weakened.

THEOREM 1.17. *Let us assume that F is a Peano map and that* $K \subset \text{Dom}(F)$ *is compact. If there exists a viable solution* $x(\cdot)$ *such that*

$$\inf_{t>0} \frac{1}{t} \int_0^t \|x'(\tau)\| d\tau = 0$$

then there exists a viable equilibrium.

When K is a compact viability domain, then the convexity of either $F(K)$ *or of* K implies the existence of a viable equilibrium.

THEOREM 1.18. *Let F be a Peano map. If $K \subset \mathrm{Dom}(F)$ is a compact viability domain and if $F(K)$ is convex, then there exists a viable equilibrium.*

The following theorem holds.

THEOREM 1.19. *Let $F: X \rightsquigarrow X$ be an upper semicontinuous set-valued map with closed convex images. If $K \subset X$ is a convex compact viability domain of F, then it contains an equilibrium of F.*[20]

This theorem is equivalent to the Brouwer Fixed Point Theorem,[21] as well as many other statements of nonlinear analysis, such as the Kakutani-Fan fixed point theorem and the versatile and efficient *Ky* Fan Inequality. It states that if K is a compact convex subset and if $\varphi: X \times X \to \mathbf{R}$ is a function satisfying

$$\begin{cases} \text{(i)} & \forall y \in K, \quad x \to \varphi(x, y) \text{ is lower semicontinuous} \\ \text{(ii)} & \forall x \in K, \quad y \to \varphi(x, y) \text{ is concave} \\ \text{(iii)} & \forall y \in K, \quad \varphi(y, y) \leqq 0 \end{cases}$$

then, there exists $\bar{x} \in K$, such that for all $y \in K$, $\varphi(\bar{x}, y) \leqq 0$.

2. Regulation of control systems.

2.1. Viable control systems. We now translate the viability theorems into the language of control theory. From now on, we introduce the state space X, the constraint space Y, the control space Z and a *feedback set-valued map* $U: X \rightsquigarrow Z$ associating with any state x the (possibly empty) subset $U(x)$ of feasible controls when the state of the system is x. In other words, we assume that the available *controls of the system are required to obey constraints which may depend upon the state.*

The dynamics of the system are further described by a (single-valued) map $f: \mathrm{Graph}(U) \to X$ which assigns to each state-control pair $(x, u) \in \mathrm{Graph}(U)$ the *velocity* $f(x, u)$ of the state.

Hence the set $F(x) := \{f(x, u)\}_{u \in U(x)}$ is the set of available velocities to the system when its state is x.

We shall assume from now on that $f(x, u) := c(x) + g(x)u$, where $g(x) \in \mathscr{L}(Z, X)$ are linear operators, i.e., that the system is affine with respect to the control. We also consider the case when the viability domain $K := h^{-1}(M)$ is defined by more explicit constraints through a map h from X to the constraint space Y:

The evolution of the system (U, f) is governed by the differential inclusion

$$(2) \qquad \begin{cases} \text{(i)} & \text{for almost all } t, \quad x'(t) = f(x(t), u(t)) \\ \text{(ii)} & \text{where } u(t) \in U(x(t)). \end{cases}$$

When $U(x) = Z$ for all $x \in K$ and when $M = \{0\}$, we recognize the traditional control systems.

The regulation map R_K is the set-valued map defined by:

$$R_K(x) := \{u \in U(x) \mid h'(x)g(x)u \in T_M(h(x)) - h'(x)c(x)\}.$$

---

[20] Actually, this theorem remains true for any Hausdorff locally convex topological vector-space and in particular, weak topologies.

[21] See [4, Chap. II] for a proof based on differential geometry.

THEOREM 2.1. *Let us assume that the dynamics satisfy*

$$(3)\quad \begin{cases} \text{(i)} & \forall (x,u)\in \text{Graph}\,(U),\quad f(x,u) =: c(x)+g(x)u \\ \text{(ii)} & \text{Graph}\,(U) \text{ is closed and the images of } U \text{ are convex} \\ \text{(iii)} & c\colon \text{Dom}\,(U)\to X \text{ is continuous} \\ \text{(iv)} & g\colon \text{Dom}\,(U)\to \mathscr{L}(Z,X) \text{ is continuous and bounded} \\ \text{(v)} & c \text{ and } U \text{ have linear growth} \end{cases}$$

and that the constraints verify

$$\begin{cases} \text{(i)} & M \text{ is a closed sleek subset of } Y \\ \text{(ii)} & h \text{ is a } \mathscr{C}^1\text{-map from } X \text{ to } Y \\ \text{(iii)} & \forall x\in K := h^{-1}(M),\quad Y := \text{Im}\,(h'(x)) - T_M(h(x)) \\ \text{(iv)} & \forall x\in h^{-1}(M),\quad \exists u\in U(x) \text{ such that} \\ & h'(x)g(x)u\in T_M(h(x)) - h'(x)c(x). \end{cases}$$

Then K is a viability domain of the control system and the viable solutions are regulated through the regulation law

$$(4)\qquad \text{for almost all } t,\quad u(t)\in R_K(x(t))$$

and the regulation map R_K has compact nonempty convex values.

When U is assumed to be lower semicontinuous, an additional uniform transversality condition implies that the regulation map is also lower semicontinuous.

We can also characterize viability domain through a dual formulation.

PROPOSITION 2.2. *We posit the assumptions of Theorem* 2.1. *Then $K := h^{-1}(M)$ is a viability domain if and only if*

$$\begin{cases} \forall (x,q)\in \text{Graph}\,(N_M), \\ d_M(x,q) := \inf_{u\in U(x)}\langle q, h'(x)g(x)u + h'(x)c(x)\rangle \leqq 0. \end{cases}$$

For instance, this condition holds true when the following abstract Walras law holds true:

$$\begin{cases} \text{(i)} & Z = Y, \qquad U(x)\subset N_M(x) \\ \text{(ii)} & \forall_q \in N_M(x),\quad \langle q, h'(x)c(x) + h'(x)g(x)q\rangle \leqq 0. \end{cases}$$

Example. Let us mention that the calculus on the contingent cones can be transferred to a calculus of regulation maps. For instance, a quite common type of viability constraints are of the form $K := L\cap h^{-1}(M)$ where we assume that

$$\begin{cases} \text{(i)} & K\subset X \text{ and } M\subset Y \text{ are sleek} \\ \text{(ii)} & h \text{ is a } \mathscr{C}^1\text{-map from } X \text{ to } Y \\ \text{(iii)} & \forall x\in K := L\cap h^{-1}(M),\quad Y = h'(x)T_K(x) - T_M(h(x)). \end{cases}$$

Indeed, K is the inverse image of the product $L\times M$ by the map $I\times h$ from X to $X\times Y$.

This is a particular case of a more general situation when both X, Y, and Z are product spaces. It may then be convenient to provide once and for all the explicit formulas of the regulation map when this is the case. Let us assume namely that

$$(5)\quad \begin{cases} \text{(i)} & X := \prod_{i=1}^n X_i & \\ \text{(ii)} & Y := \prod_{j=1}^m Y_j & M := \prod_{j=1}^m M_j \\ \text{(iii)} & Z := \prod_{k=1}^l Z_k & U(x) := \prod_{k=1}^l U_k(x) \end{cases}$$

and that

$$(6)\qquad \begin{cases} \text{(i)} & \forall x \in X, \quad g_i(x) := \sum_{k=1}^{l} g_i^k u_k, \\ \text{(ii)} & c(x) := (c_1(x), \cdots, c_n(x)),\ g(x) := (g_1(x), \cdots, g_n(x)) \\ \text{(iii)} & \forall x \in X, \quad h_j(x) := \sum_{i=1}^{n} h_j^i(x_i). \end{cases}$$

Therefore, K is the intersection of the subsets K_j defined by:

$$(7)\qquad K_j := \left\{ x \in X \;\middle|\; \sum_{i=1}^{n} h_j^i(x_i) \in M_j \right\}.$$

Let us introduce the matrix $B(x) := h'(x)g(x)$ of operators

$$B_j^k = \sum_{i=1}^{n} h_j^{i'}(x) g_i^k \in \mathscr{L}(U_k, Y_j)$$

and the vector $b(x) := h'(x)c(x)$ of components

$$b_j(x) = \sum_{i=1}^{n} h_j^{i'}(x) c_i(x).$$

Then the regulation map R_K is defined by

$$(8)\qquad \begin{cases} \text{(i)} & R_K(x) = \cap_{j=1}^{m} R_{K_j}(x) \quad \text{where} \\ \text{(ii)} & R_{K_j}(x) = \{u = (u_1, \cdots, u_l) \in \prod_{k=1}^{l} U_k(x) \quad \text{such that} \\ & \sum_{k=1}^{l} B_j^k u_k \in T_{M_j}(\sum_{i=1}^{n} h_j^i(x_i) - b_j(x))\} \end{cases}$$

and has compact values. If it is strict, then K is a viability domain of the system, and thus, for any initial state $x_0 \in K$, there exist one *viable* solution $x_i(\cdot)$ on $[0, \infty[$ starting at x_0 to the system of differential equations

$$\forall i, \quad x_i'(t) = c_i(x(t)) + \sum_{k=1}^{l} g_i^k(x(t)) u_k(t)$$

and *open loop* controls regulating this viable solution $x(\cdot)$ in the sense that the regulation law

$$\forall j, \quad \text{for almost all } t, \quad u(t) \in R_{K_j}(x(t)).$$

We shall say that the regulation map is *decoupled* if

$$Z = Y \quad \text{and} \quad \forall j \neq k, \quad B_j^k = 0.$$

In this case, each partial viability domain K_j is regulated by the ith component of the control in the sense that

$$R_{K_j}(x) = \left\{ u_j \in U_j(x) \middle| B_j^j u_j \in T_{M_j}\left(\sum_{i=1}^{n} h_j^i(x_i) - b_j(x) \right) \right\}.$$

2.2. Closed-loop controls and slow solutions. Viable solutions to the control system (2) are regulated by the viable controls whose evolution is governed by the regulation law (4). Continuous single-valued selections r_K of the regulation map R_K are *viable closed-loop controls*, since the Viability Theorem states that the differential equation

$$x'(t) = f(x(t), r_K(x(t)))$$

enjoys the viability property.

Indeed, by construction, K is a viability domain of the single-valued map $x \in K \mapsto f(x, r_K(x))$.

So, we have to investigate under which assumptions there exists a continuous selection of the regulation map: the answer is given by Michael's Theorem: *Let R be*

a lower semicontinuous set-valued map with closed convex values from a compact space X to a Banach space Y. It does have a continuous selection (see [8, Chap. 9], for instance).

Hence, we obtain the existence of viable continuous closed-loop controls.

PROPOSITION 2.3. *We posit the assumptions of Theorem* 2.1. *If* $R_K(\cdot)$ *is lower semicontinuous, the control system can regulate viable solutions in K by continuous closed loop controls.*

This result is not useful in practice, since Michael's selection theorem does not provide constructive ways to find those continuous closed-loop controls. Therefore, we are tempted to use explicit selections of the regulation map R_K, such as the minimal selection R_K^o defined by

$$R_K^o(x) := \{u \in R_K(x) \mid \|u\| = \min_{y \in R_K(x)} \|y\|\}. \tag{9}$$

It is continuous only when R is continuous with closed convex images. Unfortunately, there is no hope to have, in general, continuous regulation maps R_K (as soon as we have inequalities constraints). Hence this minimal selection is not necessarily continuous when the regulation map is only lower semicontinuous. But we can still prove that by taking the minimal selection R_K^o, the differential equation

$$x'(t) = f(x(t), R_K^o(x(t))) \tag{10}$$

does enjoy the viability property.

DEFINITION 2.4. The solutions to the differential equation (10) are called slow viable solutions to the control system (2).

THEOREM 2.5. *We posit the assumptions of Theorem* 2.1. *If* $R_K(\cdot)$ *is lower semicontinuous, then the control system has slow viable solutions.*

The reason why this theorem holds true is that the minimal selection is obtained through a "strict convex" selection procedure defined in the following way:

DEFINITION 2.6 (selection procedure). *A selection procedure of a set-valued map* $R : X \rightsquigarrow Y$ is a set-valued map $S_R : X \rightsquigarrow Y$

$$\begin{cases} \text{(i)} & \forall x \in \text{Dom}(R), \quad S(R(x)) := S_R(x) \cap R(x) \neq \emptyset \\ \text{(ii)} & \text{the graph of } S_R \text{ is closed} \end{cases}$$

and the set-valued map $S(R) : x \rightsquigarrow S(R(x))$ is called the *selection* of R.

It is said *convex-valued* or simply, *convex* if its values are convex and *strict* if moreover

$$\forall x \in \text{Dom}(R), \quad S_R(x) \cap R(x) = \{s(R(x))\} \text{ is a singleton.}$$

We can easily provide such examples of selection procedures through optimization, thanks to the Maximum Theorem.

PROPOSITION 2.7. *Let us assume that a set-valued map* $R : X \rightsquigarrow Y$ *is lower semicontinuous with compact values. Let* $V : \text{Graph}(R) \mapsto \mathbf{R}$ *be continuous. Then the set-valued map* S_R *defined by*:

$$S_R(x) := \{y \in Y \mid V(x, y) \leqq \inf_{y' \in R(x)} V(x, y')\}$$

is a selection procedure of R. Consequently, if the graph of R is also closed, so is the graph of the selection $S(R)$ *equal to*:

$$S(R(x)) = \{y \in R(x) \mid V(x, y) \leqq \inf_{y' \in R(x)} V(x, y')\}$$

or through game theoretical methods, as in the following proposition.

PROPOSITION 2.8. *Let us assume that a set-valued map $R : X \rightsquigarrow Y$ is lower semicontinuous with convex compact values. Let $\varphi : X \times Y \times Y \to \mathbf{R}$ satisfy*

$$\begin{cases} \text{(i)} & \varphi(y, y') \text{ is lower semicontinuous} \\ \text{(ii)} & \forall (x, y) \in X \times Y, \quad y' \mapsto \varphi(x, y, y') \text{ is concave} \\ \text{(iii)} & \forall (x, y) \in X \times Y, \quad \varphi(x, y, y) \leqq 0. \end{cases}$$

Then the map S_R associated with φ by the relation

$$S_R(x) := \{y \in Y \mid \sup_{y' \in R(x)} \varphi(x, y, y') \leqq 0\}$$

is a selection procedure of R. If R is also closed, so is the selection map $x \mapsto S(R(x))$.

This is the fact that the minimal selection is obtained through a strict convex selection procedure which matters. So, Theorem 2.5 can be extended to any strict convex selection of the regulation map R_K.

THEOREM 2.9. *We posit the assumptions of Theorem* 2.1. *Let S_{R_K} be a strict convex selection of the regulation map R_K. Then the single-valued selection $s(R_K)$ defined by*

$$\forall x \in K, \quad s(R_K(x)) := R_K(x) \cap S_{R_K}(x)$$

is a viable closed loop control.

Strictness of the selection procedure is needed only to obtain single-valued closed loop controls. Otherwise, the proof of the above theorem provides the existence of "selected" regulation laws, associated to selections of the regulation map.

THEOREM 2.10. *We posit the assumptions of Theorem* 2.1. *Let S_{R_K} be a convex selection of the regulation map R_K. Then, for any initial state $x_0 \in K$, there exist a viable solution starting at x_0 and a viable control to the control system* (2) *which are regulated by the selection $S(R_K)$ of the regulation map R_K, in the sense that*

$$\text{for almost all } t \geqq 0, \quad u(t) \in S(R_K)(x(t)) := R_K(x(t)) \cap S_{R_K}(x(t)).$$

2.3. Smooth solutions. There are many reasons for looking for "smooth viable controls," which are absolutely continuous instead of being measurable. Too much oscillation of the controls can damage them, for instance. Also, as we stated in the introduction, we need to differentiate viable open loop controls to implement the "inertia principle."

We can obtain smooth viable solutions by setting a bound to the growth to the evolution of controls. For that purpose, we shall associate to this control system and to any nonnegative continuous function $u \to \varphi(x, u)$ with linear growth[22] the system of differential inclusions

$$(11) \qquad \begin{cases} \text{(i)} & x'(t) = f(x(t), u(t)) \\ \text{(ii)} & u'(t) \in \varphi(x(t), u(t))B. \end{cases}$$

We observe that any solution $(x(\cdot), u(\cdot))$ to the system of differential inclusions (11) which is viable in Graph (U) is a smooth solution to the control system (2). This property is a viability requirement: the state-control pair has to be viable in the graph of U. Hence, we need to study the contingent cone to the graph of a map, which leads us to the concept of contingent derivative of a set-valued map.

[22] Which can be a constant c, or the function $c\|u\|$, or the function $(x, u) \to c(\|u\| + \|x\| + 1)$.

2.4. Contingent derivatives.

DEFINITION 2.11. We introduce the *contingent derivative* $DF(x, y)$ of a set-valued map $F: X \rightsquigarrow Y$, defined by

$$\text{Graph}\,(DF(x, y)) := T_{\text{Graph}(F)}(x, y). \tag{12}$$

We shall say that F is *sleek* at $(x, y) \in \text{Graph}\,(F)$ if and only if

$$(x', y') \rightsquigarrow \text{Graph}\,(DF)(x', y') \text{ is lower semicontinuous at } (x, y)$$

and it is *sleek* if it is sleek at every point of its graph.

Naturally, the contingent derivative is a closed convex process whenever F is sleek at (x, y).

When $F := f$ is single-valued, we set $Df(x) := Df(x, f(x))$.

If f is differentiable around a point $x \in K$, then *the contingent derivative of the restriction is the restriction of the derivative to the contingent cone*:

$$D(f|_K)(x) = D(f|_K)(x, f(x)) = f'(x)|_{T_K(x)}$$

Actually, this follows from the following useful proposition.

PROPOSITION 2.12. *Let f be a differentiable operator from an open subset $\Omega \subset X$ to Y, $M: X \rightsquigarrow Y$ be a set-valued map and $L \subset XM \subset Y$ be a closed subset. Let $F: X \rightsquigarrow Y$ be the set-valued map defined by*:

$$F(x) := \begin{cases} f(x) - M(x) & \text{when } x \in L \\ \varnothing & \text{when } x \notin L. \end{cases}$$

Let (x, y) belong to the graph of F.

Assume that either L or M is sleek. Then its contingent derivative is equal to

$$DF(x, y)(u) := \begin{cases} f'(x)u - DM(x, f(x) - y) & \text{when } u \in T_L(x) \\ \varnothing & \text{when } u \notin T_L(x). \end{cases}$$

Another familiar instance of set-valued maps is the inverse of a set-valued map F (or even of a noninjective single-valued map). We can easily compute its contingent derivative because *a contingent derivative of the inverse of a set-valued map F is the inverse of the contingent derivative*:

$$D(F^{-1})(y, x) = DF(x, y)^{-1}.$$

These contingent derivatives are characterized by adequate limits of difference quotients.

PROPOSITION 2.13. *Let $(x, y) \in \text{Graph}\,(F)$ belong to the graph of a set-valued map $F: X \rightsquigarrow Y$. Then*

$$\begin{cases} v \text{ belongs to } DF(x, y)(u) \text{ if and only if} \\ \liminf\limits_{h \to 0+, u' \to u} d\left(v, \dfrac{F(x + hu') - y}{h}\right) = 0 \end{cases} \tag{13}$$

They enjoy chain rule formulas and the Inverse-Function Theorem can be extended to set-valued maps, as shown in the following Theorem.

THEOREM 2.14 (Inverse Function Theorem). *Let us consider a closed set-valued map $F: x \rightsquigarrow Y$ and a solution x_0 to the inclusion $F(x_0) \ni y_0$. Let us assume that F is sleek at (x_0, y_0). If $DF(x_0, y_0)$ is surjective, then y_0 belongs to the interior of the image of F and F is pseudo-Lipschitz in the sense that there exist a positive constant λ and neighborhoods $\mathcal{U}$ of x and $\mathcal{V}$ of y such that*

$$\forall y_1, y_2 \in \mathcal{V}, \quad F^{-1}(y_1) \cap \mathcal{U} \subset F^{-1}(y_2) + \lambda \|y_1 - y_2\|_Y B_X.$$

In particular, by taking $F := f|_K$ and $F := f$, we obtain the constrained Inverse Function Theorem, as follows.

THEOREM 2.15 (Constrained Inverse Function). *Let us consider a (single-valued) continuous map $f: X \mapsto Y$, a closed subset $K \subset X$ and an element x_0 of K.*

We assume that f is continuously differentiable at x_0, that K is sleek at x_0 and that $f'(x_0)T_K(x_0) = Y$. Then $f(x_0)$ belongs to the interior of $f(K)$ and the set-valued map $y \rightsquigarrow f^{-1}(y) \cap K$ is pseudo-Lipschitz around $(f(x_0), x_0)$.

If $K := X$ is the whole space and if $f'(x_0)$ is surjective, we infer that the set-valued map $y \rightsquigarrow f^{-1}(y)$ is pseudo-Lipschitz around $(f(x_0), x_0)$.

2.5. Regularity Theorem. We thus deduce the following Regularity Theorem.

THEOREM 2.16. *Let us assume that the control system (2) satisfies*

$$\text{(14)} \qquad \begin{cases} \text{(i)} & \text{Graph}(U) \text{ is closed} \\ \text{(ii)} & f \text{ is continuous and has linear growth.} \end{cases}$$

Then for any initial state $x_0 \in \text{Dom}(U)$ and any initial control $u_0 \in U(x_0)$, there exists a smooth state-control solution $(x(\cdot), u(\cdot))$ to the control system (2) starting at (x_0, u_0) if and only if the set-valued map U satisfies

$$\forall (x, u) \in \text{Graph}(U), \quad DU(x, u)(f(x, u)) \cap \varphi(x, u)B \neq \emptyset.$$

The assumption of the above theorem is too strong, since it requires that it is satisfied for all controls u of $U(x)$ (so that we have a solution for every initial control chosen in $U(x_0)$). We may very well be content with the existence of a smooth solution for only some initial control in $U(x_0)$.

So, we can relax the problem by looking for the largest closed set-valued feedback map contained in U in which we can find the initial state-controls yielding smooth viable solutions to the control system. This amounts to studying the viability kernels of Graph (U) for the system of differential inclusions [11].

DEFINITION 2.17 (φ-growth regulation map). Let us consider the control system (2). We shall denote by $R^\varphi := R^\varphi_U$ the set-valued map whose graph is the viability kernel of Graph (U) for the system of differential inclusions (11). We shall call it the *φ-growth regulation map* to the control system (2). If $\varphi \equiv 0$, we shall say that R^0_U is the *punctuated regulation map.*

We thus deduce from Theorem 1.15 the following result on the existence of smooth viable solutions.

THEOREM 2.18. *Let us assume that the control system (2) satisfies*

$$\text{(15)} \qquad \begin{cases} \text{(i)} & \text{Graph}(U) \text{ is closed,} \\ \text{(ii)} & f \text{ is continuous and has linear growth.} \end{cases}$$

Then for any initial state $x_0 \in \text{Dom}(R^\varphi)$ and any initial control $u_0 \in R^\varphi(x_0)$, there exists a smooth state-control solution $(x(\cdot), u(\cdot))$ to the control system (2) starting at (x_0, u_0), where the solution $x(\cdot)$ is regulated by a control $u(\cdot)$ starting at u_0 through the φ-regulation law:

$$\text{(16)} \qquad \forall t \geqq 0, \quad u(t) \in R^\varphi(x(t)).$$

Remark. We observe that the graph of R^φ_U is also the viability kernel of the graph of the regulation map R_U and that the regulation maps R^φ are increasing with φ.

It will be interesting to relate the states and the controls which provide zero velocities.

DEFINITION 2.19 (punctuated equilibrium). We associate with any control u its *viability niche* $N^\varphi(u)$, which is the (possibly empty) closed subset of states $x \in \text{Dom}(R^\varphi)$

such that $0 \in DR^{\varphi}(x, u)(f(x, u))$. When $\varphi \equiv 0$, the viability niche $N^0(u)$ is called the *viability cell* of u. A control u is called a *punctuated equilibrium* if and only if its viability cell is not empty.

We remark at once that the set-valued map $u \rightsquigarrow N^0(u)$ is the inverse of $x \rightsquigarrow R^0(x)$ and that when $\varphi_1 \geqq \varphi_2 \geqq 0$, we have

$$N^0(u) \subset N^{\varphi_1}(u) \subset N^{\varphi_2}(u).$$

The case when the growth φ is equal to 0 is particularly interesting, because it determines areas where the evolution of the control is constant: The viability cell of a control u is the viability kernel of $U^{-1}(u)$ for the differential equation $x'(t) = f(x(t), u)$ parametrized by the constant control u. Naturally, *when the viability cell of a punctuated equilibrium is reduced to a point, this point is an equilibrium.*

2.6. Heavy solutions. Let us consider a control system (U, f) which has a nontrivial φ-growth regulation map R^{φ}_U for some nonnegative function φ.

PROPOSITION 2.20. *The smooth viable state-control pairs* $(x(\cdot), u(\cdot))$ *to the control system* (2) *are also solutions to the system of differential inclusions*

$$\begin{cases} \text{(i)} & x'(t) = f(x(t), u(t)) \\ \text{(ii)} & u'(t) \in DR^{\varphi}_U(x(t), u(t))(f(x(t), u(t))). \end{cases} \tag{17}$$

The question arises of whether we can construct selection procedures of the control component of this system of differential inclusions. It is convenient for this purpose to introduce the following definition.

DEFINITION 2.21 (dynamical closed loops). We shall say that a selection g of the contingent derivative of the φ-regulation map R^{φ}_U in the direction $f(x, u)$ defined by

$$\forall (x, u) \in \text{Graph}\,(R^{\varphi}_U), \quad g(x, u) \in DR^{\varphi}_U(x, u)(f(x, u)) \tag{18}$$

is a *dynamical closed loop.*

The system of differential equations

$$\begin{cases} \text{(i)} & x'(t) = f(x(t), u(t)) \\ \text{(ii)} & u'(t) = g(x(t), u(t)) \end{cases} \tag{19}$$

is called the associated *closed loop differential system.*

Therefore, a dynamical closed loop being given, we can select smooth viable state-control pairs as solutions to systems of ordinary differential equations.

Such solutions do exist when g is continuous (and if such is the case, they will be continuously differentiable). But they also may exist when g is no longer continuous, as we saw when we built closed-loop controls. This is the case for instance when $g(x, u)$ is the element of minimal norm in $DR^{\varphi}_U(x, u)(f(x, u))$.

In both cases, we need to assume that the right-hand side of this system is lower semicontinuous with closed convex images.

We begin by deducing from Michael's Theorem the existence of continuously differentiable viable state-control solutions.

THEOREM 2.22. *We posit the assumptions of Theorem* 2.18. *If*

$$\begin{cases} \text{(i)} & \textit{the domains of } U \textit{ and } R^{\varphi}_U \textit{ coincide} \\ \text{(ii)} & \textit{the } \varphi\textit{-regulation map } R^{\varphi}_U \textit{ is sleek} \\ \text{(iii)} & \sup_{(x, u) \in \text{Graph}\,(R^{\varphi}_U)} \|DR^{\varphi}_U(x, u)\| < +\infty \end{cases} \tag{20}$$

then there exists a continuous dynamical closed loop. The associated closed-loop differential system regulates continuously differentiable viable state-control solutions.

Since we do not know constructive ways to build continuous dynamical closed loops, we shall investigate whether some explicit dynamical closed loop provides closed loop differential systems which do possess solutions by using selection procedures of $(x, u) \rightsquigarrow DR_U^\varphi(x, u)(f(x, u))$.

The simplest example of dynamical closed loop control is the map g_φ^o associating with each state-control pair (x, u) the element of minimal norm of $DR_U^\varphi(x, u)(f(x, u))$.

DEFINITION 2.23 (heavy viable solutions). We denote by $g_\varphi^o(x, u)$ the element of minimal norm of $DR_U^\varphi(x, u)(f(x, u))$. We shall say that the solutions to the associated closed loop differential system

$$\begin{cases} \text{(i)} & x'(t) = f(x(t), u(t)) \\ \text{(ii)} & u'(t) = g_\varphi^o(x(t), u(t)) \end{cases}$$

are *heavy viable solutions* to the control system (U, f).

THEOREM 2.24 (heavy viable solutions). *We posit the assumptions of Theorem* 2.22. *Then for any initial state-control* (x_0, u_0) *in* Graph (R_U^φ), *there exists a heavy viable solution to the control system* (2).

Remark. Let $(x(\cdot), u(\cdot))$ be a heavy viable solution to the control system. We observe that if for some t_1, the solution enters the viability niche $N^\varphi(u(t_1))$; the control $u(t)$ remains equal to $u(t_1)$ as long as $x(t)$ remains in the viability niche $N^\varphi(u(t_1))$. Since a viability niche is not necessarily a viability domain, the solution may leave it.

If for some $t_f > 0$, $u(t_f)$ is a punctuated equilibrium, then $u(t) = u_{t_f}$ for all $t \geqq t_f$ and $x(t)$ *remains in the viability cell* $N^0(u(t_f))$ *for all* $t \geqq t_f$.

The reason why this theorem holds true is that the minimal selection is obtained through the strict selection procedure defined in (2.6), which is convex. This is this fact that matters. So, Theorem 2.24 can be extended to any strict convex selection of the set-valued map $DR_U^\varphi(x, u)(f(x, u))$.

For simplicity, we set

$$G_\varphi(x, u) := DR_U^\varphi(x, y)(f(x, u)).$$

THEOREM 2.25. *We posit the assumptions of Theorem* 2.18. *Let* S_{G_φ} *be a strict convex selection of the set-valued map* G_φ. *Then, for any initial state* $(x_0, u_0) \in$ graph (U), *there exists a viable state-control solution starting at* (x_0, u_0) *to the associated closed loop differential system*

$$\begin{cases} \text{(i)} & x'(t) = f(x(t), u(t)) \\ \text{(ii)} & u'(t) = s(DR_U^\varphi(x(t), u(t))(f(x(t), u(t))) \\ & \qquad := G_\varphi(x(t), u(t)) \cap S_{G_\varphi}(x(t), u(t)). \end{cases}$$

Strictness of the selection procedure of the set-valued map G_φ is needed only to obtain closed-loop systems of differential equations. Otherwise, the proof of the above theorem provides the existence of solutions to closed loop systems of "smaller" differential inclusions.

THEOREM 2.26. *We posit the assumptions of Theorem* 2.18. *Let* S_{G_φ} *be a convex selection of the set-valued map* G_φ. *Then, for any initial state* $(x_0, u_0) \in$ graph (U), *there exists a viable state-control solution starting at* (x_0, u_0) *to the associated closed loop system of differential inclusions*

$$\text{(21)} \qquad \begin{cases} \text{(i)} & x'(t) = f(x(t), u(t)) \\ \text{(ii)} & u'(t) \in S(DR_U^\varphi(x(t), u(t))f(x(t), u(t))) \\ & \qquad := G_\varphi(x(t), u(t)) \cap S_{G_\varphi}(x(t), u(t)). \end{cases}$$

2.7. Example: One-dimensional affine system. Let us illustrate the above considerations by the simplest dynamical economic model (one commodity, one consumer). See Fig. 3.

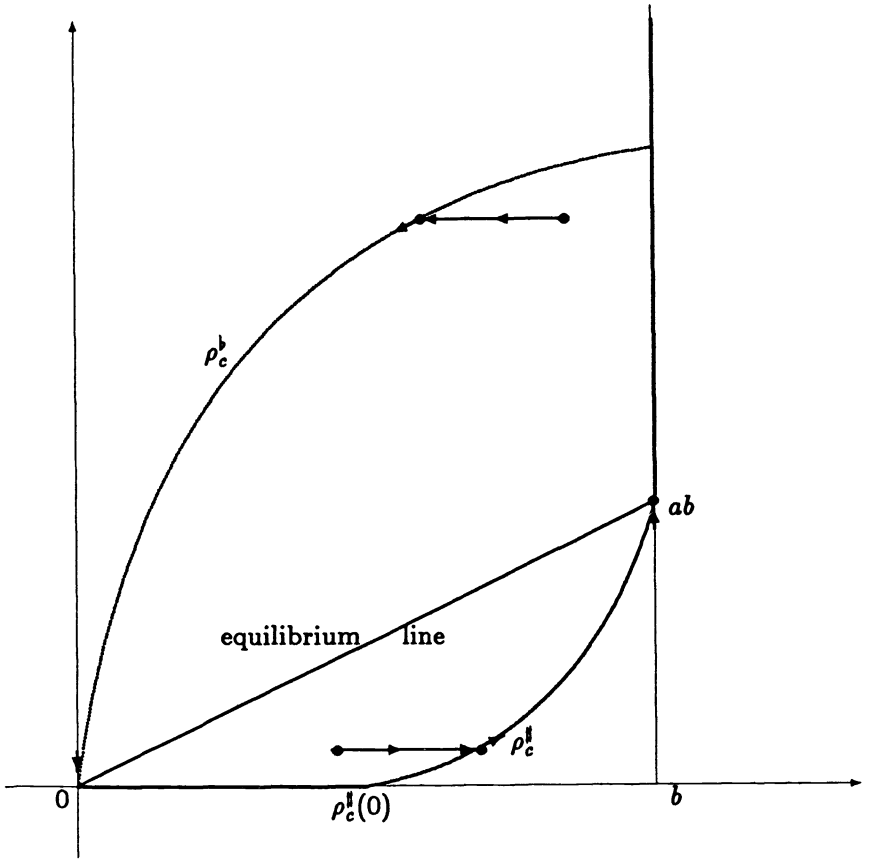


FIG. 3. *Evolution of a heavy solution.*

Let $K := [0, b]$ the subset of scarce commodities. Assume that the consumption rate of a consumer is equal to $a > 0$, so that, without any further restriction, its exponential consumption will leave the viability subset $[0, b]$. Hence its consumption is slowed down by a price which is used as a control. In summary, the evolution of its consumption is governed by the control system

$$\text{for almost all } t \geqq 0, \quad x'(t) = ax(t) - u(t), \quad \text{where } u(t) \geqq 0$$

subject to the constraints for all $t \geqq 0$, $x(t) \in [0, b]$.

The a priori feedback map U is defined by $U(x) := \mathbf{R}_+$. Hence the regulation map is given by the formula

$$R_K(0) = \{0\}, \quad R_K(x) = \mathbf{R}_+ \quad \text{when} \quad x \in]0, b[\quad \text{and} \quad R_K(b) = [ab, +\infty[.$$

Its graph is not closed, and its closure is the graph of U, equal to $[0, b] \times \mathbf{R}_+$.

We see at once that the viable equilibria of the system range over the *equilibrium line* $u = ax$. Viability is guaranteed each time that the price $u(t)$ is chosen in $R(x(t))$, i.e., $u = 0$ when $x = 0$ (and thus, the system cannot leave the equilibrium because negative prices are not allowed "to start" the system) and $u \geqq ab$ when $x = b$, so that the price is large enough to stop or decrease consumption.

Assume that the system obeys the inertia principle: *it keeps the price constant as long as it works.* Take for instance $x_0 > 0$ and $u_0 \in [0, ax_0[$. Then the consumption increases[23] and when it reaches the boundary b of the interval, the system must switch very quickly to a velocity large enough to slow down the consumption fast enough for the solution to remain in the interval $[0, b]$.

But there is a bound to growth of prices (and inflation rates), so that we should set a bound[24] on price velocities: $|u'(t)| \leqq c$. We shall associate with such a bound a "last warning" to modify the price: there is a level of consumption after which it will be impossible to slow down the consumption with a velocity smaller than or equal to c to forbid it to increase beyond the boundary b. We thus consider the c-bounded state-control solutions, which are the solutions to the system

$$(22) \qquad \begin{cases} \text{(i)} & \text{for almost all } t \geqq 0, \quad x'(t) = ax(t) - u(t) \\ \text{(ii)} & \text{and } -c \leqq u'(t) \leqq c \end{cases}$$

which are viable in Graph (U).

We introduce the functions $\rho^\#$ and $\rho^\flat$ defined on $[0, \infty[$ by

$$\begin{cases} \text{(i)} & \rho_c^\#(u) := (c/a^2)(e^{-au/c} - 1 + (a/c)u) \approx u^2/2c \\ \text{(ii)} & \rho_c^\#(u) := -ce^{a(u-ab)/c}/a^2 + u/a + c/a^2 \end{cases}$$

and the functions $r^\#$ and $r^\flat$ defined on $[0, b]$ by

$$\begin{cases} \text{(i)} & r^\flat(x) = u \quad \text{if and only if } u = \rho_c^\flat(x) \\ \text{(ii)} & r^\#(x) = 0 \quad \text{if } x \in [0, \rho_c^\#(0)] \ (\rho^\#(0) = (c/a^2)(1 - e^{+a^2 b/c})) \\ \text{(iii)} & r^\#(x) = u \quad \text{if and only if } u = \rho^\#(x) \text{ when } x \in [\rho_c^\#(0), b]. \end{cases}$$

PROPOSITION 2.27. *The c-bounded growth regulation map of system* (22) *is defined by*

$$(23) \qquad \forall x \in [0, b], \quad R^c(x) = [r^\#(x), r^\flat(x)].$$

Let us build the heavy solutions. We shall investigate the cases when the initial control u_0 is below or above the equilibrium line.

Consider the case when $x_0 > 0$ and the price $u_0 \in [r^\#(x_0), ax_0[$. Since we want to choose the price velocity with minimal norm, we take $u'(t) = 0$[25] as long as the solution $x(\cdot)$ to the differential equation $x' = ax - u_0$ yields a consumption $x(t) < \rho_c^\#(u_0)$. When for some time t_1, the consumption $x(t_1) = \rho_c^\#(u_0)$, it has to be slowed down. Indeed, otherwise $(x(t_1 + \varepsilon), u_0)$ will be below the curve $\rho_c^\#$ and in this case, any solution starting from this situation will eventually cease to be viable. Therefore, prices should increase to slow down the consumption growth. The idea is to take the smallest velocity u' such that the vector $(x'(t_1), u')$ takes the state inside the graph of R^c: they are the velocities $u' \geqq x'(t_1)/\rho_c^\#(u_0)$. By construction, it is achieved by the velocity of $x^\#(\cdot)$, which is the highest one allowed to increase prices. Therefore, by taking

$$x(t) := x^\#(t) := e^{a(t-t_1)}(x(t_1) - u_0/a - c/a^2) + c(t - t_1)/a + u_0/a + c/a^2$$

and $u(t) := u_0 + c(t - t_1)$ for $t \in [t_1, t_1 + (ab - u_0)/c]$, we get a solution which ranges over the curve $x^\#(t) = \rho_c^\#(u^\#(t))$. This is a heavy solution because, for the same reason as above, the smallest velocity of the price (which is unique along this curve) is chosen.

---

[23] It is equal to $(e^{at}(ax_0 - u_0) + u_0/a)$.

[24] We take $\varphi(x, u) \equiv c$.

[25] And realize in this case the dream of economists, which, despite the teachings of history, are looking for constant prices and commodities.

According to the above differential equation, we see that $x(t)$ increases to b where it arrives with velocity 0 and the price increases linearly until it arrives at the equilibrium price ab. Since (b, ab) is an equilibrium, the heavy solution stays there: we take $x(t) \equiv b$ and $u(t) \equiv 0$ when $t \geqq t_1 + u_0/c$. So we have built a viable solution starting from (x_0, u_0).

Consider now the case when $u_0 \in [ax_0, r^\flat(x_0)]$, where we follow the same construction of the heavy viable solution. We start by taking $u'(t) = 0$, and thus, $u(t) = u_0$, as long as the solution $x(\cdot)$ to the differential equation $x' = ax - u_0$, which decreases, satisfies $x(t) > \rho_c^\flat(u_0)$. Then, when $x(t_1) = \rho_c^\flat(u_0)$ for some t_1, we take

$$x(t) = x^\flat(t) := e^a(t - t_1)(x(t_1) - u_0/a + c/a^2) - c(t - t_1)/a + u_0/a - c/a^2$$

and $u(t) := u_0 - c(t - t_1)$ for $t \in [t_1, t_1 + u_0/c]$ in order to avoid leaving the viability kernel. Finally, for $t \geqq t_1 + u_0/c$, we take $x(t) \equiv 0$ and $u(t) \equiv 0$.

Remark. We observe that for any $x \in]0, b[$,

$$\lim_{c \to 0+} r^\flat(x) = \lim_{c \to 0+} r^\#(x) = ax, \quad \lim_{c \to \infty} r^\#(x) = 0, \quad \text{and} \quad \lim_{c \to \infty} r^\flat(x) = +\infty.$$

In other words, the graph of R^c starts from the equilibrium line when $c = 0$ and converges in some sense to the graph of U when $c \to +\infty$.

3. Lyapunov and energy functions. We consider a differential inclusion (1) and a time-dependent function $w(\cdot)$ defined as a solution to the differential equation

$$w'(t) = -\varphi(w(t)) \tag{24}$$

where $\varphi : \mathbf{R}_+ \mapsto \mathbf{R}$ is a given continuous function with linear growth. This function φ is used as a parameter in what follows.

Our problem is to characterize functions enjoying the φ-Lyapunov property, i.e., nonnegative extended functions $V : X \to \mathbf{R}_+ \cup \{+\infty\}$ (such that $\mathrm{Dom}(V) \subset \mathrm{Dom}(F)$) satisfying

$$\forall t \geqq 0, \quad V(x(t)) \leqq w(t), \quad w(0) = V(x(0)) \tag{25}$$

along at least one solution $x(\cdot)$ to the differential inclusions (1) and (24). Since this condition amounts to saying that the epigraph of V enjoys the viability property for the differential inclusion

$$(x'(t), w'(t)) \in G(x(t), w(t)) \quad \text{where } G(x, w) := F(x) \times \{-\varphi(w)\}$$

we can apply the Viability Theorem. This allows us to use lower semicontinuous instead of differentiable functions among the candidates to satisfy this Lyapunov property. We have to translate the fact that $\mathscr{E}p(V)$ is a viability domain of G, and for that purpose, to study the contingent cones to the epigraphs. This leads us to the concept of contingent epiderivatives of an extended function.

3.1. Epiderivatives of real-valued functions. Let us then consider an extended real-valued function $V : X \mapsto \mathbf{R} \cup \{+\infty\}$ whose domain

$$\mathrm{Dom}(V) := \{x \in X \mid V(x) < +\infty\}$$

is not empty. (Such a function is said to be *proper* in convex and nonsmooth analysis. We shall rather say that it is *nontrivial* to avoid confusion with proper maps.) They are characterized by their epigraphs

$$\mathscr{E}p(V) := \{(x, \lambda) \in X \times \mathbf{R} \mid V(x) \leqq \lambda\}.$$

An extended function V is lower semicontinuous (respectively, convex) if and only if its epigraph is closed (respectively, convex).

The main examples of extended functions are the *indicators* ψ_K *of subsets* K defined by $\psi_K(x) := 0$ if $x \in K$ and $+\infty$ if not. They are lower semicontinuous if and only if the subsets are closed and convex if and only if the subsets are convex. One can regard the sum $V + \psi_K$ as the restriction of V to K.

DEFINITION 3.1 (epiderivatives). Let $V : X \mapsto \mathbf{R} \cup \{+\infty\}$ be a nontrivial extended real-valued function and x belong to its domain. We shall say that the function $D_\uparrow(V)(x)$ from X to $\mathbf{R} \cup \{+\infty\} \cup \{-\infty\}$ defined by

$$D_\uparrow(V)(x)(u) := \liminf_{h \to 0+, u' \to u} (V(x + hu') - V(x))/h$$

is the *contingent epiderivative* of V at x in the direction u.

The function is said to be *contingently epidifferentiable* if its contingent epiderivative is *nontrivial* in the sense that it never takes the value $-\infty$ and has at least one finite value. It is said to be *episleek* if and only if its epigraph is sleek.

Naturally, the contingent epiderivative coincides with the directional derivative $\langle V'(x), u \rangle$ when V is Gâteaux differentiable.

If V is continuously differentiable around a point $x \in K$, then the *contingent epiderivative of the restriction is the restriction of the derivative to the contingent cone*:

$$D_\uparrow(V|_K)(x)(u) := \begin{cases} \langle V'(x), u \rangle & \text{if } u \in T_K(x) \\ +\infty & \text{if not.} \end{cases}$$

We observe the following proposition.

PROPOSITION 3.2. *Let* $V : \mapsto \mathbf{R} \cup \{+\infty\}$ *be an extended function and* x *belong to its domain. Then the contingent cone to the epigraph of* V *at* $(x, V(x))$ *is the epigraph of the contingent epiderivative of* V *at* x: $\mathcal{E}pD_\uparrow V(x) = T_{\mathcal{E}p(V)}(x, V(x))$. *Furthermore,*

$$\{D_\uparrow V(x)(u), D_\uparrow V(x)(u)\} \subset DV(x)(u) \subset [D_\uparrow V(x)(u), D_\downarrow V(x)(u)].$$

These subsets are equal when V *is episleek.*

3.2. Lyapunov functions.

DEFINITION 3.3 (Lyapunov functions). We shall say that a nonnegative contingently epidifferentiable extended function V is a *Lyapunov function* of F associated with a function $\varphi(\cdot) : \mathbf{R}_+ \mapsto \mathbf{R}$ if and only if V is a solution to the "contingent[26] Hamilton-Jacobi inequalities"

$$\forall x \in \mathrm{Dom}(V), \quad \inf_{v \in F(x)} D_\uparrow V(x)(v) + \varphi(V(x)) \leqq 0. \tag{26}$$

THEOREM 3.4. *Let* V *be a nonnegative contingently epidifferentiable lower semicontinuous extended function and* $F : X \rightsquigarrow X$ *be a Peano map. Then* V *is a Lyapunov function of* F *associated with* $\phi(\cdot)$ *if and only if for any initial state* $x_0 \in \mathrm{Dom}(V)$, *there exist solutions* $x(\cdot)$ *to* (1) *and* $w(\cdot)$ *to* (24) *satisfying property* (25).

Example. W-monotone set-valued maps. Let $W : X \to \mathbf{R}_+ \cup \{+\infty\}$ be a nonnegative extended function. We say that a set-valued map F is *W-montone* (*with respect to* ϕ) if

$$\forall x, y, \quad \forall u \in F(x), v \in F(y), D_\uparrow W(x - y)(v - u) + \phi(W(x - y)) \leqq 0. \tag{27}$$

We obtain, for instance, the following consequence.

COROLLARY 3.5. *Let* W *be a nonnegative contingently epidifferentiable extended lower semicontinuous function and* $F : X \rightsquigarrow X$ *be a Peano map such that* $-F$ *is W-monotone*

[26] We refer to [55], [56], [61] and the references of these papers for a thorough study of contingent Hamilton-Jacobi equations arising from optimal control and comparison with *viscosity solutions* introduced by Crandall and Lions (see [41]).

with respect to some ϕ. Let $\bar{x}$ be an equilibrium of F (i.e., $0 \in F(\bar{x})$). Then, for any initial state x_0, there exist solutions $x(\cdot)$ and $w(\cdot)$ satisfying

$$\forall t \geqq 0, \quad W(x(t) - \bar{x}) \leqq w(t).$$

In particular, for $W(z) := \frac{1}{2}\|z\|^2$, we find the usual concept of monotonicity (with respect to ϕ):

$$\forall x, y, \quad \forall u \in F(x), v \in F(y), \quad \langle u - v, x - y \rangle \geqq \phi(\tfrac{1}{2}\|x - y\|^2).$$

We can reformulate the Viability Theorem in the following way.

COROLLARY 3.6. *Let $F: X \rightsquigarrow X$ be a Peano map. A closed subset K enjoys the viability property if and only if its indicator Ψ_K is a solution to the contingent equation*

$$\inf_{v \in F(x)} D_\uparrow \Psi_K(x)(v) = 0.$$

Let us also introduce attractors in the following definition.

DEFINITION 3.7. We shall say that a closed subset K is an "attractor" of order $\alpha \geqq 0$ if and only if for any $x_0 \in \mathrm{Dom}(F)$, there exists at least one solution $x(\cdot)$ to the differential inclusion (1) such that

$$\forall t \geqq 0, \quad d_K(x(t)) \leqq d_K(x_0) e^{-\alpha t}. \tag{28}$$

In the following corollary we can recognize attractors by checking whether the distance function to K is a Lyapunov function.

COROLLARY 3.8. *Assume that F is a Peano map. Then a closed subset $K \subset \mathrm{Dom}(F)$ is an attractor if and only if the function $d_K(\cdot)$ is a solution to the contingent inequalities*:

$$\forall x \in \mathrm{Dom}(F), \quad \inf_{v \in F(x)} D_\uparrow d_K(x)(v) + \alpha d_K(x) \leqq 0.$$

Remark. With an extended nonnegative function V, we can associate affine functions $w \to aw - b$ for which V is a solution to the contingent Hamilton-Jacobi inequalities (26).

For that purpose, we consider the convex function b defined by

$$b(a) := \sup_{x \in \mathrm{Dom}(F)} \left(\inf_{v \in F(x)} D_\uparrow V(x)(v) + aV(x) \right).$$

Then it is clear that V is a solution to the contingent Hamilton-Jacobi inequalities

$$\forall x \in \mathrm{Dom}(F), \quad \inf_{v \in F(x)} D_\uparrow V(x)(v) + aV(x) - b(a) \leqq 0.$$

Therefore, we deduce that there exists a solution to the differential inclusion satisfying

$$\forall t \geqq 0, \quad V(x(t)) \leqq \left(V(x_0) - \frac{b(a)}{a} \right) e^{-at} + \frac{b(a)}{a}.$$

A reasonable choice of a is the largest of the minimizers of $a \in]0, \infty[\to \max(0, b(a)/a)$, for which $V(x(t))$ decreases as fast as possible to the smallest level set $V^{-1}(]-\infty, (b/a)])$ of V.

The functions φ and $U: X \to \mathbf{R} \cup \{+\infty\}$ being given, we can construct the smallest lower semicontinuous Lyapunov function larger than or equal to U, i.e., the smallest nonnegative lower semicontinuous solution U_φ to the contingent Hamilton-Jacobi inequalities (26) larger than or equal to U. Its epigraph is the viability kernel of the epigraph of U, as seen in the following theorem.

THEOREM 3.9. *Let us consider a Peano map* $F: X \rightsquigarrow X$, *a continuous function* $\varphi: \mathbf{R}_+ \to \mathbf{R}$ *with linear growth and a proper nonnegative extended function* U *such that* $\mathrm{Dom}(U) \subset \mathrm{Dom}(F)$. *Then there exists a smallest nonnegative lower semicontinuous solution* $U_\varphi: \mathrm{Dom}(F) \mapsto \mathbf{R} \cup \{+\infty\}$ *to the contingent Hamilton-Jacobi inequalities* (26) *larger than or equal to* U (*which can be the constant* $+\infty$), *which enjoys the property*:

$$\forall x_0 \in \mathrm{Dom}(U_\varphi), \quad \begin{array}{l}\text{there exist solutions to (1) and (24)} \\ \text{satisfying } \forall t \geqq 0, \quad U(x(t)) \leqq U_\varphi(x(t)) \leqq w(t).\end{array}$$

Let us single out the following consequence.

COROLLARY 3.10. *We posit the assumptions of Theorem* 3.9. *For all* $a \geqq 0$, *there exists a smallest lower semicontinuous function* $d_{M_a}: X \to \mathbf{R} \cup \{+\infty\}$ *larger than or equal to* d_M *such that*

$$\forall x_0 \in \mathrm{Dom}(d_{M_a}), \quad \begin{array}{l}\text{there exists a solution } x(\cdot) \text{ to (1) such that} \\ d_M(x(t)) \leqq d_{M_a}(x_0) e^{-at}.\end{array}$$

Therefore, we can regard the subsets $\mathrm{Dom}(d_{M_a})$ *as the "basins" of exponential attraction of* M.

3.3. Asymptotic observability of differential inclusions. Let us consider a set-valued map F from $X := \mathbf{R}^n$ to X and an observation map h from X to $Y := \mathbf{R}^p$. We "observe" the evolution

$$\forall t \geqq 0, \quad y(t) := h(x(t))$$

of an unknown solution $x(\cdot)$ to the differential inclusion (1).

The problem is to "simulate asymptotically" at least an unknown state $x(\cdot)$ by a solution $z(\cdot)$ to a control system where the control is the observation of the state

$$z'(t) = g(z(t), y(t)). \tag{29}$$

We shall measure the asymptotic behavior of the error $x(\cdot) - z(\cdot)$ through a nonnegative lower semicontinuous extended function $U: X \mapsto \mathbf{R} \cup \{+\infty\}$ and through a function $w(\cdot)$ from $[0, +\infty]$ to $\mathbf{R}_+$ by inequalities

$$\forall t \geqq 0, \quad U(x(t) - z(t)) \leqq w(t) \tag{30}$$

where $w(\cdot)$ is a solution to differential equation (24).

DEFINITION 3.11. Let F, h, φ and U be given. We say that the dynamical system F observed through h is *stabilizable* by g with respect to U and φ if

$$\forall x, z, \quad \inf_{v \in F(x)} D_\uparrow U(x-z)(v - g(z, h(x))) \leqq -\varphi(U(x-z)).$$

THEOREM 3.12. *We assume that* F *is a Peano map, that* g, h, *and* φ *are continuous with linear growth and that* $U: X \mapsto \mathbf{R}_+ \cup \{+\infty\}$ *is contingently epidifferentiable, lower semicontinuous and episleek. If the dynamical system* F *observed through* h *is stabilizable by* g, *then for any initial state* x_0 *and* z_0, *there exist solutions* $x(\cdot)$ *to* (1), $z(\cdot)$ *to* (29) *and* $w(\cdot)$ *to* (24) *starting at* x_0, z_0 *and* $U(x_0 - z_0)$, *respectively, and satisfying inequalities* (30).

We now have to construct stabilizing maps g in various situations.

We begin by providing a first class of examples using (U, φ)-monotone maps.

PROPOSITION 3.13. *Let us assume that* U, φ, f, *and* h *being given, we can find a continuous map* $c: Y \to X$ *such that*

$$\text{the map } x \mapsto c(h(x)) - F(x) \text{ is } (U, \varphi)\text{-monotone.}$$

Then for any continuous selection f of F, the single-valued map

$$g(z, y) := f(z) - c(h(z)) + c(y)$$

stabilizes F through h with respect to U and φ.

The problem now is to recognize whether there exist functions U and φ and a map c which makes the set-valued map $c \circ h - F$ to be (U, φ)-monotone.

More generally, let us introduce the set-valued map H defined by

$$H(z, x) := \{v \mid \inf_{u \in F(x)} D_\uparrow U(x - z)(u - v) + \varphi(U(x - z)) \leqq 0\}.$$

The general problem of stabilizing F through h amounts to finding selections g of the set-valued map G defined by

$$\forall (z, y), \quad G(z, y) = \bigcap_{h(x)=y} H(z, x)$$

since by construction, such selections are stabilizing f through h. When G is lower semicontinuous with closed convex values, Michael's Theorem guarantees the existence of a continuous selection. Hence, in this case, we can stabilize F, at least in theory, since Michael's Theorem is not constructive.

3.4. Lyapunov preorders. Let us consider more generally a preorder $\geqq$ and look for solutions $x(\cdot)$ of differential inclusion (1) which do not decrease in the sense that

$$\forall t \geqq s \geqq 0, \quad x(t) \geqq x(s).$$

For that purpose, it is useful to characterize a preorder by the set-valued map P defined[27] by

$$\forall x, \quad P(x) := \{y \mid y \geqq x\}$$

the graph of which is the graph of the preorder.

PROPOSITION 3.14. *Let F be a Peano map and P be a preorder with closed graph whose domain is contained in the domain of F.*

The following statements are equivalent:

$$(31) \quad \begin{cases} \text{(i)} & \forall x \in \mathrm{Dom}\,(P), \quad F(x) \cap T_{P(x)}(x) \neq \varnothing \\ \text{(ii)} & \forall (x, y) \in \mathrm{Graph}\,(P), \quad F(y) \cap DP(x, y)(0) \neq \varnothing \\ \text{(iii)} & \forall x_0 \in \mathrm{Dom}\,(P), \quad \exists x(\cdot) \in \mathscr{S}(x_0) \text{ such that} \\ & \forall t \geqq s \geqq 0, \quad x(t) \geqq x(s). \end{cases}$$

The same type of proof yields results dealing with the comparison of solutions to two differential inclusions, as shown in the following proposition.

PROPOSITION 3.15. *Let $F : X \rightsquigarrow X$ and $G : X \rightsquigarrow X$ be two Peano maps and a preorder P with closed graph whose graph is contained in* $\mathrm{Dom}\,(F) \cap \mathrm{Dom}\,(G)$.

Then the following statements are equivalent:

$$(32) \quad \begin{cases} \text{(i)} & \forall (x, y) \in \mathrm{Graph}\,(P), \quad G(y) \cap DP(x, y)(F(x)) \neq \varnothing \\ \text{(ii)} & \forall x_0 \in \mathrm{Dom}\,(P), \exists x(\cdot) \in \mathscr{S}_F(x_0) \quad \text{and} \quad y(\cdot) \in \mathscr{S}_G(x_0) \text{ such that} \\ & \forall t \geqq 0, \quad y(t) \geqq x(t). \end{cases}$$

[27] When the preorder is defined by q functions V_i, the set-valued map P associates with any x the subset $P(x) := \{y \mid V_i(y) \leqq V_i(x) (i = 1, \cdots, q)\}$. Its graph is closed if and only if V_i are continuous on their domains.

3.5. Fuzzy differential inclusions. Using differential inclusions for representing uncertainty can be criticized on the ground that it gives velocities of the system at state x the same "likeness" to be chosen. Is there a possibility to discriminate among velocities and to choose among the viable ones those which are somewhat better?

To answer this problem we suggest replacing the usual subset of velocities in the right-hand side of the differential inclusion by a "fuzzy set" of velocities. Fuzzy sets are represented by "membership functions" χ taking their values in the interval $[0, 1]$, the membership functions of usual subsets being their characteristic functions, taking their values in $\{0, 1\}$. Here, we characterize subsets by their indicators ψ_K, taking their values in $\{0, +\infty\}$, so that membership functions of "fuzzy subsets" are extended functions $V: X \mapsto \mathbf{R} \cup \{+\infty\}$, which measure, in some sense, the cost of belonging to the fuzzy subset.

DEFINITION 3.16. We shall regard an extended nonnegative function $U: X \mapsto \mathbf{R}_+ \cup \{+\infty\}$ as a *fuzzy set.* Its *domain* is the domain of U, i.e., the set of elements x such that $U(x)$ is finite.

We shall say that the fuzzy set U is *closed* (respectively, *convex*) if the extended function U is lower semicontinuous (respectively, convex).

Hence the membership function of the empty set is the constant function equal to $+\infty$.

DEFINITION 3.17. We shall say that a set-valued map $\mathbf{U}: X \rightsquigarrow Y$ associating to any $x \in X$ a fuzzy subset $U(x)$ of Y is a *fuzzy set-valued map.* Its *graph* is the fuzzy subset of $X \times Y$ associated to the extended nonnegative function $(x, y) \to U(x, y) := U(x)(y)$.

A fuzzy set-valued map $\mathbf{U}$ is said to be *closed* if and only if its graph is closed, i.e., if its membership function is lower semicontinuous. Its values are closed (respectively, convex) if and only if the fuzzy subset $U(x)$ is closed (respectively, convex). It has linear growth if and only if, for some positive constant c,

$$U(x, v) < +\infty \Rightarrow \|v\| \leqq c(\|x\| + 1).$$

By using indicators, we can reformulate differential inclusion (1) as

$$\text{for almost all } t, \quad \psi_{F(x(t))}(x'(t)) < +\infty.$$

Then we are led to define "fuzzy dynamics" of a system by a fuzzy set-valued map $\mathbf{U}$ associating to any $x \in X$ a fuzzy set $U(x)$ of velocities $\{v | U(x, v) < +\infty\}$. In this case, we can write the associated *fuzzy differential inclusion* in the form

$$\text{for almost all } t \geqq 0, \quad U(x(t), x'(t)) < +\infty \tag{33}$$

or, equivalently, in the form

$$\text{for almost all } t \geqq 0, \quad (x(t), x'(t)) \in \text{Graph}\,(\mathbf{U})$$

which is a fuzzy subset instead of a usual subset.

We shall say that a subset $K \subset \text{Dom}\,(\mathbf{U})$ is a viability domain of the fuzzy set-valued map $\mathbf{U}$ if and only if

$$\forall x \in K, \quad \exists v \in T_K(x) \text{ such that } U(x, v) < +\infty.$$

When the fuzzy set-valued map $\mathbf{U}$ is continuous, we can select a viable solution to the fuzzy differential inclusion (33) which is *sharpest*, in the sense that the cost of its velocity's membership is minimal:

$$\text{for almost all } t, \quad U(x(t), x'(t)) = \inf_{v \in T_K(x(t))} U(x(t), v). \tag{34}$$

THEOREM 3.18. *Let us consider a nontrivial fuzzy set-valued map* $\mathbf{U}$ *from a finite-dimensional vector-space* X *to itself. Let us assume that it is a Peano map. We assume moreover that the restriction of the membership function* U *to its domain (the graph of* $\mathbf{U}$*) is continuous and that the viability domain* K *is sleek.*

Then there exists a sharpest viable solution to the differential inclusion (33) *(i.e., which satisfies condition* (34)*).*

3.6. Tracking solutions to a differential inclusion. Let us consider a differential inclusion (1) where $F: X \rightsquigarrow X$ is a Peano map.

Let us introduce now an observation map $H: X \rightsquigarrow Y$ from X to another finite-dimensional vector-space Y.

We shall in some sense "project" the differential inclusion (1) to a differential equation on the observation space Y described by a set-valued map G

$$\text{for almost all } t \geqq 0, \quad y'(t) \in G(y(t)) \tag{35}$$

in order to "track" (or "filter") a solution $(x(\cdot))$ to differential inclusion (1) in the following sense:

$$\begin{cases} \forall x_0 \in \mathrm{Dom}\,(F) \text{ and } y_0 \in H(x_0), \quad \text{there exist} \\ \text{solutions } x(\cdot) \text{ and } y(\cdot) \text{ to (1) and (35)} \\ \text{such that } \forall t \geqq 0, \quad y(t) \in H(x(t)). \end{cases} \tag{36}$$

This property may be called the *tracking property.* (System (35) is called an *exosystem* by Byrnes and Isidori).

PROPOSITION 3.19. *Let us consider a closed set-valued map* H *from* X *to* Y*. Let us assume that* $F: X \rightsquigarrow X$ *and* $G: Y \rightsquigarrow Y$ *are nontrivial Peano maps and that the graph of* H *is closed. Then tracking property* (36) *holds true if and only if*

$$\forall x \in \mathrm{Im}\,(H), \quad \forall x \in H^{-1}(y), \quad F(x) \cap DH(x, y)^{-1}(G(y)) \neq \emptyset. \tag{37}$$

It follows obviously from Viability Theorem 1.8, because the above condition amounts to saying that

$$\forall (x, y) \in \mathrm{Graph}\,(H), \quad (F(x) \times G(y)) \cap T_{\mathrm{Graph}(H)}(x, y) \neq \emptyset$$

i.e., that the graph of H is a viability domain of the set-valued map $F \times G$.

Example. Energy maps. The simplest dynamics are obtained when $G \equiv 0$: in this case, each subset $H^{-1}(y)$ is a viability domain, because, for any $y \in \mathrm{Im}\,(H)$ and $x_0 \in H^{-1}(y)$, there exists a solution $x(\cdot)$ such that $x(t) \in H^{-1}(y_0)$ for all $t \geqq 0$. We shall say that such a set-valued map H is an *energy map* of F.

When $H \equiv V$ is a single-valued map from X to $Y := \mathbf{R}$, and when V is sleek, we deduce that V is an energy function if and only if

$$\forall x, \quad \exists u \in F(x) \quad \text{such that } D_{\uparrow} V(x)(u) \leqq 0 \leqq D_{\downarrow} V(x)(u)$$

because in this case[28] $DV(x)(u) = [D_{\uparrow} V(x)(u), D_{\downarrow} V(x)(u)]$.

The question arises to find such energy maps of a set-valued map F with closed graph. We deduce from Theorem 1.15 an answer to this question.

[28] When V is differentiable and $F := f$ is single-valued, we find the classical characterization

$$\langle V'(x), f(x) \rangle = \sum_{i=1}^{n} \frac{\partial V}{\partial x_i}(x) f_i(x) = 0.$$

PROPOSITION 3.20. *Let $F: X \rightsquigarrow X$ be a Peano map. Then there exists a largest closed energy map $H_0: X \rightsquigarrow Y$ of F, a solution to the inclusion*

$$\forall x \in \mathrm{Dom}\,(H), \quad \forall y \in H(x), \quad DH(x, y)(F(x)) \ni 0.$$

The graph of H_0 is the viability kernel of the set-valued map $(x, y) \rightsquigarrow F(x) \times \{0\}$.

Remark. More generally, the behavior of observations of some solutions to the differential inclusion $x' \in F(x)$ will be given by the behavior of solutions to differential equations $y' = g(y)$ whenever g is a smooth selection of

$$g(y) \in \bigcap_{x \in H^{-1}(y)} \bigcap_{v \in F(x)} DH(x, y)(v).$$

In the case when the differential equation $y' = g(y)$ has a unique solution $r(t)y_0$ starting from y_0, the solution $x(\cdot)$ satisfies the condition

$$\forall t \geqq 0, \quad x(t) \in H^{-1}(r(t)y(0)), \quad x(0) \in H^{-1}(y(0)).$$

When g is a linear operator $G \in \mathscr{L}(Y, Y)$, it can be written

$$\forall t \geqq 0, \quad x(t) \in H^{-1}(e^{Gt}y(0)), \quad x(0) \in H^{-1}(y(0)).$$

Such maps g are selections of the map $G_H: Y \rightsquigarrow Y$ defined by

$$G_H(y) := \bigcap_{y \in H^{-1}(x)} \bigcup_{v \in F(x)} DH(x, y)(v).$$

This set-valued map measures, so to speak, a degree of disorder of the system, because the larger the images of G_H, the more observed dynamics g tracking an evolution of the differential inclusion.

An adequate observation of a differential equation or inclusion is a map H, set-valued or single-valued, which provides a "small" set-valued map G_H.

When $H \equiv h$ is a single-valued differentiable map, then the map G_H can be written

$$G_H(y) := \bigcap_{h(x)=y} h'(x)F(x)$$

and a single-valued map g is a selection of G_H if and only if

$$\forall x \in \mathrm{Dom}\,(H), \quad 0 \in h'(x)F(x) - g(h(x)).$$

Example. Let us consider the case of *descriptor systems*

$$Ex'(t) = Ax(t) + Bu(t)$$

which we want to observe by $H \in \mathscr{L}(X, Y)$ through the linear equation

$$y'(t) = Gy(t)$$

where $G \in \mathscr{L}(Y, Y)$. We introduce the matrices (A, GH) from X to $X \times Y$ and

$$\begin{pmatrix} E & B \\ H & 0 \end{pmatrix} \quad \text{from } X \times Z \text{ to } X \times Y.$$

We observe that the system enjoys the tracking property (36) if and only if

$$\mathrm{Im}\,(A, GH) \subset \mathrm{Im}\begin{pmatrix} E & B \\ H & 0 \end{pmatrix}.$$

In this case, the velocities $x'(t)$ and the controls $u(t)$ are supplied by the linear system

$$\begin{cases} Ex'(t) - Bu(t) = Ax(t) \\ Hx'(t) = GHx(t) \end{cases}$$

which can be solved by linear algebraic formulas.

3.7. Decentralizing a control system. Let $H: X \rightsquigarrow Y$ be an observation map. We consider two control systems:

$$\begin{cases} \text{(i)} & \text{for almost all } t \geqq 0, \quad x'(t) = f(x(t), u(t)) \\ \text{(ii)} & \text{where } u(t) \in U(x(t)) \end{cases} \tag{38}$$

and

$$\begin{cases} \text{(i)} & \text{for almost all } t \geqq 0, \quad y'(t) = g(y(t), v(t)) \\ \text{(ii)} & \text{where } v(t) \in V(y(t)) \end{cases} \tag{39}$$

on the state and observation spaces, respectively, where $U: X \rightsquigarrow Z_X$ and $V: Y \rightsquigarrow Z_X$ map X and Y to the control spaces Z_X and Z_Y and where $f: \text{Graph}(U) \mapsto X$ and $g: \text{Graph}(V) \mapsto Y$.

We introduce the set-valued maps $R_H(x, y): Z_Y \rightsquigarrow Z_X$ defined by

$$R_H(x, y; v) = \begin{cases} \{u \in U(x) \mid f(x, u) \in DH(x, y)^{-1}(g(y, v))\} & \text{if } v \in V(y) \\ \varnothing & \text{if } v \notin V(y). \end{cases}$$

COROLLARY 3.21. *Assume that the set-valued maps U and V are Peano maps and that the maps f and g are continuous, affine with respect to the controls and with linear growth. The two control systems enjoy the tracking property* (36) *for any initial condition* $(x_0, y_0) \in \text{Graph}(H)$ *if and only if*

$$\forall (x, y) \in \text{Graph}(H), \quad \text{Graph}(R_H(x, y)) \neq \varnothing.$$

Then the system is regulated by the regulation law

$$\textit{for almost all } t \geqq 0, \quad u(t) \in R_H(x(t), y(t); v(t)).$$

When $H \equiv h$ is single-valued and differentiable, and when we set $f(x, u) := c(x) + C(x)u$ and $g(y, v) := d(y) + D(y)v$ where $C(x)$ and $D(y)$ are linear operators, we obtain the formula

$$R_H(x; v) := U(x) \cap (h'(x)C(x))^{-1}(d(h(x)) - h'(x)c(x) - D(h(x))v).$$

Example. Decentralization of a control system. We assume that the viability set of the control system (38) is defined by constraints of the form $K := L \cap A^{-1}(M)$ where

$$\begin{cases} \text{(i)} & K \subset X \text{ and } M \subset Y \text{ are sleek} \\ \text{(ii)} & A \text{ is a } \mathscr{C}^1\text{-map from } X \text{ to } Y \\ \text{(iii)} & \forall x \in K := L \cap A^{-1}(M), \quad Y = A'(x)T_K(x) - T_M(A(x)). \end{cases} \tag{40}$$

We shall "decentralize" this system by coupling it to the control system (39) defined on the space of constraints.

We associate with these two systems decoupled viability constraints

$$\begin{cases} \text{(i)} & \forall t \geqq 0, \quad x(t) \in L \\ \text{(ii)} & \forall t \geqq 0, \quad A(x(t)) = y(t) \\ \text{(iii)} & \forall t \geqq 0, \quad y(t) \in M. \end{cases} \tag{41}$$

It is obvious that the *state component* $x(\cdot)$ of any solution $(x(\cdot), y(\cdot))$ to the system ((38), (39)) satisfying viability constraints (41) is a solution to the initial control system (38) viable in K.

On the other hand, viable solutions to the decentralized system (38) can be obtained in two steps. First, $y(\cdot)$ is a viable solution in M to the control system (39) and then, the solution $x(\cdot)$ is a solution of the control system (38) satisfying the viability constraints

$$(42) \qquad \begin{cases} \text{(i)} & \forall t \geqq 0, \quad x(t) \in L \\ \text{(ii)} & \forall t \geqq 0, \quad A(x(t)) = y(t) \end{cases}$$

which do not involve the subset $M \subset Y$ of constraints anymore.

We know that the regulation map of the initial system is defined by

$$R_K(x) = \{u \in U(x) \cap T_L(x) \,|\, A'(x)f(x, u) \in T_M(A(x))\}.$$

The regulation map of the projected control system (39) is defined by

$$R_M(y) = \{v \in V(x) \,|\, g(y, v) \in T_M(y)\}.$$

This decentralization problem is a particular case of the observation problem for the set-valued map H defined by

$$H(x) := \begin{cases} A(x) & \text{if } x \in L \text{ and } A(x) \in M \\ \varnothing & \text{if not} \end{cases}$$

whose contingent derivative is equal under assumptions (40) to

$$DH(x)(u) := \begin{cases} A'(x)u & \text{if } u \in T_L(x) \text{ and } A'(x)u \in T_M(A(x)) \\ \varnothing & \text{if not.} \end{cases}$$

We introduce now the set-valued map R_H which is equal to

$$R_H(x; v) := \{u \in U(x) \cap T_L(x) \,|\, A'(x)f(x, u) = g(A(x), v)\}.$$

We observe that

$$\forall x \in K, \quad R_H(x; R_M(A(x))) \subset R_K(x).$$

The regulation map regulating solutions to the system ((38), (39)) satisfying viability conditions (41) is equal to $x \rightsquigarrow R_H(x, R_M(A(x)))$. Therefore, the regulation law linking the controls to the solutions are given by: for almost all $t \geqq 0$

$$\begin{cases} \text{(i)} & v(t) \in R_M(y(t)) \\ \text{(ii)} & u(t) \in R_H(x(t); v(t)). \end{cases}$$

The first law regulates the viable solutions to the control system (39) and the second the solutions to the control system (38) satisfying the viability constraints (42).

The reason why we call this decentralization is because the particular case when $X := Y^n$, when $A(x) := \sum_{i=1}^n x_i$ and when the control system (38) is

$$\forall i = 1, \cdots, n, \quad x_i'(t) = f_i(x(t), u_i(t)) \text{ where } u_i(t) \in U_i(x_i(t))$$

constrained by

$$\forall i = 1, \cdots, n, \quad x_i(t) \in L_i \text{ and } \sum_{i=1}^n x_i(t) \in M.$$

This system can be decentralized first by solving the viability problem for system (39) in the viability set M through the regulation law $v(t) \in R_M(y(t))$, and then, by

solving the n control systems through the regulation law

$$(u_1(t), \cdots, u_n(t)) \in R_H(x_1(t), \cdots, x_n(t); v(t)).$$

3.8. Decomposition property. For simplicity, we restrict ourselves here to the case when the observation map $H \equiv h := h_2 \circ h_1$ is the product of two single-valued and differentiable maps $h_1 : X \mapsto Y_1$ and $h_2 : Y_1 \mapsto Y_2$. Can we observe the evolution of a solution to a control problem (38) through $h_2 \circ h_1$ by observing it first through h_1 by a control system

$$(43) \qquad \begin{cases} \text{(i)} & \text{for almost all } t \geqq 0, \quad y_1'(t) = g_1(y_1(t), v_1(t)) \\ \text{(ii)} & \text{where } v_1(t) \in V_1(y_1(t)) \end{cases}$$

and then, observing this system through h_2. We introduce the maps R_h, R_{h_1}, and R_{h_2} defined, respectively, by

$$\begin{cases} R_h(x; v) = \{u \in U(x) \,|\, h'(x)f(x, u) = g(h(x), v) \quad \text{if } v \in V(h(x)), \\ R_{h_1}(x; v_1) = \{u \in U(x) \,|\, h_1'(x)f(x, u) = g_1(h_1(x), v_1) \quad \text{if } v_1 \in V(h_1(x)), \\ R_{h_2}(x_1; v) = \{v_1 \in V_1(x_1) \,|\, h_2'(x_1)g_1(x_1, v_1) = g(h_2(x_1), v) \quad \text{if } v \in V(h_2(x_1)), \end{cases}$$

and we see at once that

$$R_{h_1}(x; R_{h_2}(h_1(x); v)) \subset R_h(x; v).$$

Therefore, if the graph of $v \rightsquigarrow R_{h_1}(x; R_{h_2}(h_1(x); v))$ is not empty, we can recover from the evolution of a solution $y(\cdot)$ to the control system (39) a solution $y_1(\cdot)$ to the control system (43) by the tracking law

$$\text{for almost all } t, \quad v_1(t) \in R_{h_2}(y_1(t), v(t))$$

and then, a solution $x(\cdot)$ to the control system (38) by the tracking law

$$\text{for almost all } t, \quad u(t) \in R_{h_1}(x(t), v_1(t)).$$

REFERENCES[29]

[1] Z. ARTSTEIN (to appear), *Stabilizing selections of differential inclusions*, preprint.

[2] J.-P. AUBIN, H. FRANKOWSKA, AND C. OLECH (1986), *Controllability of convex processes*, SIAM J. Control Optim., 24, pp. 1192-1211.

[3] J.-P. AUBIN AND A. CELLINA (1984), *Differential Inclusions*, Springer-Verlag, Berlin, New York.

[4] J.-P- AUBIN AND I. EKELAND (1984), *Applied Nonlinear Analysis*, Wiley-Interscience, New York.

[5] J.-P. AUBIN AND H. FRANKOWSKA (1989), *Observability of systems under uncertainty*, SIAM J. Control Optim., 27, pp. 949-975.

[6] ——— (1987), *On the inverse function theorem*, J. Math. Pure Appl., 66, pp. 71-89.

[7] ——— (to appear), *Controllability and observability of control systems under uncertainty*, volume dedicated to Opial. IIASA RR.

[8] ——— (1990), *Set-Valued Analysis*, Birkhäuser, Basel.

[9] J.-P. AUBIN (to appear), *Differential games: a viability approach*, SIAM J. Control Optim.

[10] ——— (to appear), *Viability Theory*.

[11] G. BASILE AND G. MARRO (1969), *Controlled and conditional invariant subspaces in linear system theory*, J. Optim. Theory Appl., 3, pp. 396-415.

[12] D. J. BENDER AND A. J. LAUB (1987), *The linear-quadratic optimal regulator for descriptor systems*, IEEE Trans. Automat. Control, 32, pp. 672-688.

[13] L. D. BERKOVITZ (1986), *Differential games of generalized pursuit and evasion*, SIAM J. Control Optim., 24, pp. 361-373.

[29] We only provide references which are not quoted in [3].

[14] L. D. BERKOVITZ (1987), *Differential Games Without the Isaacs Condition*, Optimization Software, New York, pp. 308-336.
[15] ——— (1988), *Characterizations of the values of differential games*, Appl. Math. Optim., 17, pp. 177-183.
[16] ——— (1988), *Differential games of survival*, J. Math. Anal. Appl., 29, pp. 493-504.
[17] ——— (to appear), *A survey of recent results in differential games*, preprint.
[18] V. I. BLAGODATSIKH AND A. F. FILIPPOV (1985), *Differential inclusions and optimal control*, Trudy Mat. Inst. AN SSSR, 169, pp. 194-252.
[19] A. BLAQUIERE, F. GERARD, AND G. LEITMAN (1969), *Quantitative and Qualitative Games*, Academic Press, New York.
[20] A. BRESSAN AND G. COLOMBO (to appear), *Existence and continuous dependence for discontinuous O.D.E.'s*, Boll. Un. Mat. Ital., submitted.
[21] A. BRESSAN (to appear), *Directionally continuous selections and differential inclusions*, preprint, University of Colorado.
[22] ——— (to appear), *Dual variational methods in optimal control theory*, preprint.
[23] ——— (to appear), *On the qualitative theory of lower semicontinuous differential inclusions*, preprint, University of Colorado.
[24] ——— (to appear), *Upper and lower semicontinuous differential inclusions. A unified approach*, Controllability and Optimal Control.
[25] C. I. BYRNES AND B. D. O. ANDERSON (1984), *Output feedback and generic stabilizability*, SIAM J. Control Optim., 22, pp. 362-379.
[26] C. E. BYRNES AND A. ISIDORI (1984), *A frequency domain philosophy for nonlinear systems, with applications to stabilization and adaptive control*, 23rd IEEE Conf. Dec. Control, Las Vegas, Dec. 7-9, pp. 1569-1573.
[27] ——— (to appear), *The analysis and design of nonlinear feedback systems.* I. *Zero dynamics and global normal forms*, preprint.
[28] ——— (to appear), *The analysis and design of nonlinear feedback systems.* II. *Global stabilization of minimum phase systems*, preprint.
[29] ——— (to appear), *Feedback design from the zero dynamics point of view*, preprint.
[30] ——— (to appear), *Output regulation of nonlinear systems*, preprint.
[31] S. L. CAMPBELL AND K. D. CLARK (to appear), *Singular control problem structure and the convergence of backward differential formula*, preprint.
[32] S. L. CAMPBELL (1980), *Singular Systems of Differential Equations*, Pitman, Boston, MA.
[33] ——— (1982), *Singular Systems of Differential Equations* II, Pitman Advanced Publishing Program, Boston, MA.
[34] ——— (1986), *Consistent initial conditions for linear time varying systems*, in Frequency Domains and State Spaces Methods for Linear Systems, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam.
[35] ——— (1987), *A general form for solvable linear time varying singular systems of differential equations*, SIAM J. Math. Anal., 18, pp. 1101-1115.
[36] A. CELLINA AND A. ORNELAS (1988), *Convexity and the closure of the solution set to differential inclusions*, preprint, Scuole Internationale Superiore di Studi Avanzati.
[37] ——— (1988), *Representation of the attainable set for Lipschitzian differential inclusions*, preprint, Scuole Internationale Superiore di Studi Avanzati.
[38] A. CELLINA (to appear), *On the set solutions to Lipschitzian differential inclusions*, preprint, Scuole Internationale Superiore di Studi Avanzati.
[39] G. COLOMBO, A. FONDA, AND A. ORNELAS-GONCALVES (to appear), *Lower semicontinuous perturbations of maximal monotone differential inclusions*, preprint, Scuole Internationale Superiore di Studi Avanzati.
[40] M. CORLESS, G. LEITMANN, AND E. P. RYAN (1984), *Tracking in the presence of bounded uncertainties*, Proceedings of the Fourth International Conference on Control Theory, Sidney.
[41] M. G. CRANDALL AND P. L. LIONS (1983), *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277, pp. 1-42.
[42] K. DEIMLING AND M. R. RAO MONANA (to appear), *On solution sets of multivalued differential equations*, J. Math. Phys. Sci.
[43] K. DEIMLING (1985), *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, New York.
[44] ——— (1987), *Existence theorems for multivalued differential equations*, Proc. Intern. Sympos. Nonlinear Analysis and Application to Biomathematics, Andhra Univ., Waltair.
[45] ——— (1988), *External solution of multivalued differential equations*, Results of Mathematics, preprint.
[46] ——— (1988), *Multivalued differential equations on closed sets*, Differential and Integral Equations, 1, pp. 23-30.

[47] ——— (1989), *Extremal solutions of multivalued differential equations* II, preprint, Results in Math.
[48] M. FALCONE AND P. SAINT-PIERRE (1987), *Slow and quasi-slow solutions of differential inclusions*, J. Nonlinear Anal., T.M.A., 3, pp. 367-377.
[49] C. FOIAS, G. R. SELL, AND R. TEMAM (1986), *Inertial manifolds for nonlinear evolutionary equations*, preprint, IMA série 234.
[50] H. FRANKOWSKA (1990), *A priori estimates for operational differential inclusions*, J. Differential Equations, 84, pp. 100-128.
[51] ——— (to appear), *Hamilton-Jacobi equation: viscosity solutions and generalized gradients*, J. Math. Anal.
[52] ——— (1990), *On controllability and observability of implicit systems.*
[53] ——— (to appear), *Set-Valued Analysis and Control Theory*, Birkhäuser, Basel.
[54] ——— (1990), *Some inverse mapping theorems*, Nonlinear Analyses, Ann. Inst. H. Poincaré.
[55] ——— (1987), *L'équation d'Hamilton-Jacobi contingente*, Comptes Rendus de l'Académie des Sciences, Paris, Série 1, 304, pp. 295-298.
[56] ——— (1987), *Optimal trajectories associated to a solution of Hamilton-Jacobi equations*, IEEE, 26th CDC Conference, Los Angeles, December 9-11.
[57] ——— (1988), *Nonsmooth solutions of Hamilton-Jacobi-Bellman equations*, Proceedings of the International Conference Bellman Continuum, Antibes, France, June 13-14, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, New York.
[58] ——— (1989), *Contingent cones to reachable sets of control systems*, SIAM J. Control Optim., 27, pp. 170-198.
[59] ——— (1989), *High order inverse function theorems*, Proceedings of the Congrès Franco-Quebecois on Nonlinear Analysis and Application, Perpignan, June 1987; also in Analyse Non Linéaire, H. Attouch, J. P. Aubin, F. H. Clarke, and I. Ekeland, eds., Gauthier-Villars, 1989.
[60] ——— (1989), *Local controllability of control systems with feedbacks*, Proceedings of 25th CDC Conference, IEEE, Athens, December 10-12.
[61] ——— (1989), *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi equations*, Appl. Math. Optim., 19, pp. 291-311.
[62] ——— (1989), *Local controllability of control systems with feedbacks*, J. Optim. Theory Appl., 60, pp. 277-296.
[63] G. HADDAD (1981), *Monotone viable trajectories for functional differential inclusions*, J. Differential Equations, 42, pp. 1-24.
[64] ——— (1981), *Monotone trajectories of differential inclusions with memory*, Israel J. Math., 39, pp. 38-100.
[65] ——— (1981), *Topological properties of the set of solutions for functional differential inclusions*, Nonlinear Anal. Theory, Meth. Appl., 5, pp. 1349-1366.
[66] O. HAJEK (1975), *Pursuit Games*, Academic Press, New York.
[67] C. J. HIMMELBERG AND F. S. VAN VLECK (1986), *Existence of solutions for generalized differential equations with unbounded right-hand side*, J. Differential Equations, 61, pp. 295-320.
[68] J. HOFBAUER AND K. SIGMUND (1988), *The Theory of Evolution and Dynamical Systems*, Cambridge University Press, London Math. Soc., # 7, Cambridge.
[69] A. ISIDORI (1989), Nonlinear Control Systems, Springer-Verlag, Berlin, New York.
[70] A. ISIDORI AND C. H. MOOG (1988), *On the nonlinear equivalent of the notion of transmission zeros*, in Modelling and Adaptive Control, C. Byrnes and A. Kurzhanski, eds., LNCIS 105, Springer-Verlag, Berlin, New York, pp. 146-158.
[71] P. IVO VRKOC AND V. KRIVAN (to appear), *Absolutely continuous selections from absolutely continuous set valued map*, preprint.
[72] N. N. KRASOVSKI AND A. I. SUBBOTIN (1974), *Positional Differential Games*, Nauka, Moscow.
[73] N. N. KRASOVSKI (1986), *The Control of a Dynamic System*, Nauka, Moscow.
[74] A. KRENER, A. ISIDORI, C. GORI-GIORGI, AND S. MONACO (1981), *Nonlinear decoupling via feedback: a differential geometric approach*, IEEE Trans. Automat. Control, AC-26, pp. 331-345.
[75] ——— (1980), *Nonlinear zero distributions*, 19th IEEE Conf. Decision and Control.
[76] ——— (1983), *Linearization by output injection and nonlinear observers*, Sys. Control Lett., 3, pp. 47-52.
[77] A. J. KRENER (1985), *Nonlinear controller design via approximate normal forms*, preprint.
[78] V. KRIVAN (to appear), *Construction of population growth equations in the presence of viability constraints*, preprint.
[79] ——— (to appear), *Perturbation of viability problems*, preprint.
[80] A. B. KURZHANSKII AND T. F. FILIPPOVA (1986), *On the description of the set of viable trajectories of a differential inclusion*, Doklady AN SSSR, 289, pp. 38-41.
[81] ——— (1986), *On viable solutions for uncertain systems*, Doklady AN SSSR.

[82] A. B. KURZHANSKII (1985), *On the analytical description of the viable solutions of a controlled system*, Uspekhi Mat. Nauk, 4.

[83] ——— (1986), *On the analytical properties of viability tubes of trajectories of differential systems*, Doklady Acad. Nauk SSSR, 287, pp. 1047-1050.

[84] YU. S. LEDYAEV (1985), *Regular differential games with mixed constraints on the controls*, Proceedings of the Steklov Institute of Mathematics, 167, pp. 233-242.

[85] G. LEITMAN, B. P. RYAN, AND A. STEINBERG (1986), *Feedback control of uncertain systems: robustness with respect to neglected actuator and sensor dynamics*, Internat. J. Control, 43, pp. 1243-1356.

[86] P.-L. LIONS (1982), *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, MA.

[87] S. LOJASIEWICZ JR. (1985), *Some theorems of Scorza-Dragoni type for multifunctions with applications to the problem of existence of solutions for differential multivalued equations*, Mathematical Control Theory, pp. 625-643.

[88] D. G. LUENBERGER (1977), *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, AC-22, 1977, pp. 312-321.

[89] H. MARCHAUD (1938), *Sur les champs de demi-cônes et les equations différentielles du premier order*, Bull. Sci. Math., 62, pp. 1-38.

[90] G. MARRO (1975), *Fondamenti Di Teoria Dei Sistemi*, Patron Editore, Rome.

[91] S. MONACO AND D. NORMAND-CYROT (1988), *Zero dynamics of sampled linear systems*, Systems Control Lett.

[92] B. C. MOORE AND A. J. LAUB (1978), *Computation of supremal (A, B)-invariant and controllability subspaces*, IEEE Trans. Automat. Control, AC-23, pp. 783-792.

[93] B. E. PADEN AND S. S. SASTRY (1987), *A calculus for computing Filippov's differential inclusion with application to the variable structure control of robot manipulators*, IEEE Trans. Circuits Systems, 34, pp. 73-82.

[94] L. PANDOLFI (1981), *On the regulator problem for linear degenerate control systems*, J. Optim. Theory Appl., 33, pp. 241-254.

[95] N. H. PAVEL (1984), *Differential Equations, Flow Invariance and Applications*, Pitman Research Notes in Mathematics, 113, Pitman, Boston, MA.

[96] ——— (1987), *Nonlinear evolution operators and semigroups; applications to partial differential equations*, Lecture Notes in Mathematics, 1260, Springer-Verlag, Berlin, New York.

[97] PHAN VAN CHUONG (1985), *Equations différentielles—Un résultat d'existence de solutions pour des équations différentielles multivoques*, C. R. Acad. Sci. Paris, 301, pp. 399-402.

[98] G. S. POLOVINKIN AND G. V. SMIRNOV (1986), *Differentiation of multivalued mappings and properties of solutions of differential inclusions*, Soviet Math. Dokl., 33, pp. 662-666.

[99] M. QUINCAMPOIX (1988), *Playable differentiable games*, ILASA working paper.

[100] P. SAINT-PIERRE (to appear), *Approximation of slow solutions to differential inclusions*, preprint.

[101] E. SCHECHTER (to appear), *A survey of local existence theories for abstract nonlinear initial value problems*, Nonlinear Semigroups, Partial Differential Equations and Attractors, Springer Lecture Notes in Mathematics Series, Springer-Verlag, Berlin, New York, preprint.

[102] S. SCHWABIK (1985), *Generalized differential equations: fundamental results*, Rozpravy Ceskosloveriske Akademie ved Rada Matermatickych A. Prirodnich ved, rocnick 95, sesit 6.

[103] SHI SHUZHONG (to appear), *Nagumo type condition for partial differential inclusions*, Nonlinear Anal., T.M.A.

[104] ——— (to appear), *Théorèmes de viabilité pour les inclusions aux dérivées partielles*, preprint.

[105] ——— (to appear), *Viability theorems for a class of differential-operator inclusions*, J. Differential Equations.

[106] ——— (to appear), *Viability theory for partial differential inclusions*, Cahier de MD nx 8601, Université Paris-Dauphine.

[107] L. M. SILVERMAN (1969), *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, 14, pp. 270-276.

[108] M. STACCHETTI (to appear), *Analysis of a dynamic, decentralized exchange economy*, preprint, University of Minnesota.

[109] A. I. SUBBOTIN, N. N. SUBBOTINA, AND V. E. TRETJAKOV (1985), *Stochastic and deterministic differential inequalities*, Problems of Control and Information Theory, 14, pp. 405-419.

[110] F. TAKENS (1984), *Constrained equations: a study of implicit differential equations and their discontinuous solutions*, Dynamical Systems, Warwick, ed., Lecture Notes in Mathematics, 468, Springer-Verlag, Berlin, New York, pp. 143-233.

[111] P. TALLOS (to appear), *Viability problems for nonautonomous differential inclusions*, preprint, IIASA, Laxenburg.

[112] N. A. TCHOU (1988), *Existence of slow monotone solutions to a differential inclusion*, J. Math. Anal. Appl.

[113] M. VALADIER (1988), *Approximation Lipschitzienne par l'interieur d'une multifonction S.C.I.*, Seminaire d'Analyse Convexe, Montpellier, 1987, Expose nx 11.

[114] I. VALYI (1986), *Ellipsoidal approximations in problems of control*, IIASA working paper.

[115] PHAN VAN CHUONG (1985), *Un résultat d'existence pour des équations différentielles multivoques*, Comptes-rendus de l'Académie des Sciences, Paris, Vol. 301, pp. 399-402.

[116] A. J. VAN DES SCHAFT (1987), *On realization of nonlinear systems described by high-order differential equations*, Math. Systems Theory, 19, pp. 239-275.

[117] J. C. WILLEMS AND C. COMMAULT (1981), *Disturbance decoupling by measurement feedback with stability or pole placement*, SIAM J. Control Optim., 19, pp. 490-504.

[118] J. C. WILLEMS (1981), *Almost invariant subspaces: an approach to high gain feedback design, part* I: *Almost controlled invariant subspaces*, IEEE Trans. Automat. Control, 26, pp. 235-252.

[119] ——— (1982), *Almost invariant subspaces: An approach to high gain feedback design, part* II: *Almost conditionally invariant subspaces*, IEEE Trans. Automat. Control, 27, pp. 1071-1085.

[120] W. M. WONHAM (1985), *Linear Multivariable Control. A Geometric Approach*, Springer-Verlag, Berlin, New York.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 789-809, July 1990

002

A GENERALIZED SECOND-ORDER DERIVATIVE IN NONSMOOTH OPTIMIZATION*

R. COMINETTI† AND R. CORREA†

Abstract. In this work a new notion of generalized second-order directional derivative and generalized Hessian for nonsmooth real-valued functions is studied. The general properties of these mathematical objects are investigated together with some calculus rules that may facilitate their practical computation.

Two applications of these derivatives in optimization theory are considered: first, to obtaining necessary and sufficient second-order optimality conditions for problems with or without constraints; and second, to extending the Newton method for the minimization of a $\mathscr{C}^{1,1}$ function.

Key words. generalized second-order derivatives, nonsmooth analysis, nonsmooth optimization

AMS(MOS) subject classifications. 26A27, 26E15, 49A52, 49B27, 49D37

Introduction. The prominent role that nonsmooth analysis plays in connection with optimization theory is widely recognized, especially since the latter has natural mechanisms that generate nonsmoothness (even when starting from smooth situations): duality theory, sensitivity and stability analysis, decomposition techniques, etc.

It is therefore natural that after the achievement of a fairly complete theory of first-order generalized differentiability, in recent years interest turned toward the construction of a meaningful theory of second-order generalized differentiability (see, for instance, [1], [3], [4], [6]-[9], [11]-[15], [17]-[21], [25]-[34] and references therein) that could be used in formulating second-order conditions for optimality and eventually for constructing second-order minimization methods.

In this paper, which is basically an extended version of the results presented in [13] and [11], we describe one of the possible approaches to second-order generalized differentiability, which can be thought of as the natural second-order extension of Clarke's (first-order) derivatives [10].

More precisely, our starting point is the introduction of the generalized second-order directional derivative

$$f^{\infty}(x; u, v) = \limsup_{\substack{y \to x \\ s,t \to 0}} \frac{f(y+su+tv) - f(y+su) - f(y+tv) + f(y)}{st},$$

which turns out to be an upper semicontinuous function of x, and sublinear with respect to each direction u and v separately. This last fact permits us to introduce the generalized Hessian of f at x as the point-to-set map

$$\partial^2 f(x)(u) = \{x^* \in X^* : \langle x^*, v\rangle \leqq f^{\infty}(x; u, v),\ \forall v \in X\},$$

which is shown to be an odd fan in Ioffe's terminology, that is to say a sort of *set-valued linear map.*

At this point we should say that the previous approach bears some strong relation to the theories developed in [1], [19], and [22] as will be discussed in this paper. We should also mention that this second-order directional derivative f^{∞} recently has been considered in [26] (see also [23]), even if its study is restricted to the finite-dimensional case and no dual object as $\partial^2 f(x)$ is introduced.

* Received by the editors June 15, 1987; accepted for publication (in revised form) June 5, 1989.
† Departamento de Matemáticas, Universidad de Chile, Casilla 170-3, Correo 3, Santiago, Chile.

The paper is organized as follows. The basic definitions and general properties as well as a useful characterization of the generalized second-order derivatives are exposed in § 1. The topological properties of the so-defined derivatives are then studied in § 2 in connection with two notions of second-order Lipschitzian property.

Section 3 develops some calculus rules that facilitate the computation of the generalized second-order derivatives of a function that is built up from better behaved functions through sums, composition, and maximum.

In § 4 we prove a second-order Taylor expansion involving the generalized Hessian, and we use it to show how the convexity of a function is related to the positive semidefiniteness of its generalized Hessian.

The last two sections are concerned with the application of the generalized second-order derivatives in optimization theory. In § 5 we use them to derive necessary and sufficient second-order optimality conditions for constrained and unconstrained problems; while in § 6 we present some preliminary ideas concerning the generalized Newton method for minimizing a $\mathcal{C}^{1,1}$ function

$$x_{k+1} \in x_k - \partial^2 f(x_k)^{-1}(\nabla f(x_k)),$$

and briefly explore its application to the minimization of the augmented Lagrangian appearing in the multiplier methods for solving constrained optimization problems.

1. Generalized Hessian and second-order directional derivatives. In the sequel we will be working with real-valued mappings defined on a real Hausdorff locally convex topological vector space X (l.c.t.v.s.). We will denote by X^* the (topological) dual space of X and by $\langle \cdot, \cdot \rangle$ the canonical pairing between X and X^*, the topologies in X and X^* being compatible with the duality.

We will also consider the extended real field $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ with the usual extended operations, order, and topology familiar from convex analysis.

DEFINITION 1.1. The generalized second-order directional derivative of a function $f: X \to \mathbb{R}$ at $x \in X$ in the direction $(u, v) \in X \times X$ is defined by

$$f^\infty(x; u, v) = \limsup_{\substack{y \to x \\ t,s \to 0}} \frac{f(y+su+tv) - f(y+su) - f(y+tv) + f(y)}{st}, \tag{1}$$

and the generalized Hessian of f at x as the point-to-set mapping $\partial^2 f(x): X \rightrightarrows X^*$ given by

$$\partial^2 f(x)(u) = \{x^* \in X^* : \langle x^*, v \rangle \leqq f^\infty(x; u, v) \text{ for all } v \in X\}. \tag{2}$$

Remark. It is not difficult to see that, in the above upper limit, s and t can be taken strictly positive (or negative).

The following proposition gives the basic facts concerning f^∞ and $\partial^2 f$ for an arbitrary function f.

We recall (cf. [22]) that a point-to-set map $A: X \rightrightarrows X^*$ is called a *prefan* if it has closed and convex images, $0 \in A(0)$, and $A(tu) = tA(u)$ for all $t > 0$ and $u \in X$. The prefan A is said to be *odd* if $A(-u) = -A(u)$ for all $u \in X$, and it is called a *fan* if it has nonempty images and satisfies $A(u+v) \subset A(u) + A(v)$ for all $u, v \in X$, where the bar denotes closure.

PROPOSITION 1.2. *Let $f: X \to \mathbb{R}$ and $x \in X$. Then we have that*

(a) *The map $(u, v) \to f^\infty(u; u, v)$ is symmetric ($f^\infty(x; u, v) = f^\infty(x; v, u)$) and bisublinear (sublinear on each variable separately).*

(b) *The map* $y \to f^\infty(y; u, v)$ *is upper semicontinuous* (u.s.c.) *at* x *for every* $(u, v) \in X \times X$ *and the point-to-set map* $y \to \partial^2 f(y)(u)$ *is closed at* x *for each fixed* $u \in X$.

(c) $f^\infty(x; u, -v) = f^\infty(x; -u, v) = (-f)^\infty(x; u, v)$.

(d) $\partial^2 f(x)$ *is an odd prefan.*

Proof. (a) The symmetry of $f^\infty(x; \cdot, \cdot)$, being obvious from the definition, it suffices to show that $f^\infty(x; u, \cdot)$ is positively homogeneous and convex. If we denote

$$\Delta_f^2(y, s, t, u, v) = \frac{f(y+su+tv) - f(y+su) - f(y+tv) + f(y)}{st},$$

then the positive homogenity of $f^\infty(x; u, \cdot)$ is a direct consequence of the equality

$$\Delta_f^2(y, s, t, u, \alpha v) = \alpha \Delta_f^2(y, s, \alpha t, u, v).$$

Similarly, for every $v, w \in X$, a straightforward calculation gives

$$\Delta_f^2(y, s, t, u, v+w) = \Delta_f^2(y+tw, s, t, u, v) + \Delta_f^2(y, s, t, u, w),$$

and taking upper limits we conclude

$$f^\infty(x; u, v+w) \leqq f^\infty(x; u, v) + f^\infty(x; u, w).$$

(b) Let us show that $x \to f^\infty(x; u, v)$ is upper semicontinuous (u.s.c.). Indeed, for all $k > f^\infty(x; u, v)$ we may find an open neighborhood W of x and $\varepsilon > 0$ such that

$$\Delta_f^2(y, s, t, u, v) < k \quad \text{for all } y \in W \text{ and } 0 < |s|, |t| < \varepsilon;$$

therefore, for each $x' \in W$ we have

$$f^\infty(x'; u, v) = \limsup_{\substack{y \to x' \\ s,t \to 0}} \Delta_f^2(y, s, t, u, v) < k,$$

which shows that $f^\infty(\cdot; u, v)$ is u.s.c. at x.

To prove the closedness of $\partial^2 f(\cdot)(u)$ we just observe that for arbitrary nets $x_\alpha \to x$ and $x_\alpha^* (\in \partial^2 f(x_\alpha)(u)) \to x^*$ we have

$$\langle x_\alpha^*, v \rangle \leqq f^\infty(x_\alpha; u, v) \quad \forall v \in X,$$

so that going to the limit we obtain

$$\langle x^*, v \rangle \leqq \limsup_\alpha f^\infty(x_\alpha; u, v) \leqq f^\infty(x; u, v) \quad \forall v \in X.$$

(c) This follows directly from the equality

$$\Delta_f^2(y, s, t, u, -v) = \Delta_f^2(y, -s, -t, -u, v) = \Delta_{(-f)}^2(y, s, -t, u, v).$$

(d) Clearly, $\partial^2 f(x)(u)$ is closed and convex as an intersection of the closed half spaces $E_v = \{x^* : \langle x^*, v \rangle \leqq f^\infty(x; u, v)\}$. Also, $0 \in \partial^2 f(x)(0) = \{0\}$ since $f^\infty(x; 0, \cdot) \equiv 0$. Finally, for each $t \neq 0$ and $u \in X$ we have

$$\begin{aligned} \partial^2 f(x)(tu) &= \{x^* : \langle x^*, v \rangle \leqq f^\infty(x; tu, v) = f^\infty(x; u, tv), \ \forall v \in X\} \\ &= \{x^* : \langle x^*/t, v \rangle \leqq f^\infty(x; u, v), \ \forall v \in X\} \\ &= t \partial^2 f(x)(u), \end{aligned}$$

which proves that $\partial^2 f(x)$ is an odd prefan (for oddness take $t = -1$). □

Before illustrating the above notions with some examples, we present the next two useful formulas for the computation of f^∞.

PROPOSITION 1.3. *Let $f: X \to \mathbb{R}$ be a continuous function that admits a directional derivative $f'(y; w) = \lim_{t \downarrow 0} [f(y+tw) - f(y)]/t$ at every point $y \neq x$. Then,*

$$f^{\infty}(x; u, v) = \limsup_{\substack{y \to x \\ t \to 0}} \frac{f'(y+tu; v) - f'(y; v)}{t}, \tag{3}$$

where y and t are to be chosen so that $f'(\cdot; v)$ exists at y and $y+tu$. In other terms, $f^{\infty}(x; u, v)$ is the Clarke directional derivative (see formula (6) below) of $f'(\cdot; v)$ at x in the direction u.

Moreover, if $f'(\cdot; v)$ is continuous, then

$$f^{\infty}(x; u, v) = \limsup_{y \to x} D_+^2 f(y; u, v), \tag{4}$$

where

$$D_+^2 f(y; u, v) = \liminf_{t \downarrow 0} \frac{f'(y+tu; v) - f'(y; v)}{t}.$$

To prove this result we need two preliminary lemmas. The first one is a mean value theorem for the Dini directional derivatives of f

$$D_+ f(y; w) = \liminf_{t \downarrow 0} \frac{f(y+tw) - f(y)}{t}, \tag{5}$$

and the second is a useful characterization of the Clarke directional derivative of f

$$f^0(x; v) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+tv) - f(y)}{t}, \tag{6}$$

in terms of $D_+ f$. A straightforward application of this characterization (see (8)) allows us to derive (4) from (3).

LEMMA 1.4. *Let $f: X \to \mathbb{R}$ be continuous and $x, v \in X$, $t > 0$. Then there exists $\alpha \in]0, t[$ such that*

$$f(x+tv) - f(x) \leqq t D_+ f(x+\alpha v; v). \tag{7}$$

Moreover, if $0 < t_1 < \cdots < t_r < t$, then α may be chosen in $]0, t[\setminus \{t_1, \cdots, t_r\}$.

Proof. By introducing $h(s) = f(x+sv) - s/t[f(x+tv) - f(x)]$, (7) is equivalent to the existence of $\alpha \in]0, t[$ such that

$$h_+(\alpha) = \liminf_{t \downarrow 0} \frac{h(\alpha+t) - h(\alpha)}{t} \geqq 0.$$

Now, if $t_0 \in]0, t]$ is a global maximum of h and $t_1 \in]0, t_0[$, then we either have $h_+(t_1) \geqq 0$, in which case we are done, or $h_+(t_1) < 0$ in which case h attains a local minimum at some $\alpha \in]t_0, t_1[$ and clearly $h_+(\alpha) \geqq 0$.

Concerning the second assertion, we may use (7) repeatedly to find $\alpha_i \in]t_{i-1}, t_i[$ for $i = 1, \cdots, r+1$ (here $t_0 = 0$, $t_{r+1} = t$) such that

$$f(x+tv) - f(x) = \sum_{i=1}^{r+1} f(x+t_i v) - f(x+t_{i-1} v) \leqq \sum_{i=1}^{r+1} (t_i - t_{i-1}) D_+ f(x+\alpha_i v; v),$$

and by setting α equal to the α_i that gives the maximal $D_+ f(x+\alpha_i v; v)$ we conclude

$$f(x+tv) - f(x) \leqq \sum_{i=1}^{r+1} (t_i - t_{i-1}) D_+ f(x+\alpha v; v) = t D_+ f(x+\alpha v; v),$$

with $\alpha \in]0, t[\setminus \{t_1, \cdots, t_r\}$. □

LEMMA 1.5. *Let f, x, and v as above. Then*

$$f^0(x; v) = \limsup_{y \to x} D_+ f(y; v). \tag{8}$$

Moreover, the point $y = x$ may be ignored when taking this upper limit.

Proof. Since $D_+ f(y; v) \leqq f^0(y; v)$ and $f^0(\cdot, v)$ is u.s.c., we have

$$\limsup_{\substack{y \to x \\ y \neq x}} D_+ f(y; v) \leqq \limsup_{y \to x} D_+ f(y; v) \leqq f^0(x; v).$$

Now, if $y \in X$ and $t > 0$ we may use Lemma 1.4 to find $\alpha \in]0, t[$ with $y + \alpha u \neq x$ and

$$\frac{f(y+tu) - f(y)}{t} \leqq D_+ f(y + \alpha u; u).$$

The result follows by taking upper limits. □

We may now proceed with the proof.

Proof of Proposition 1.3. As noted previously, (4) follows from (3) and Lemma 1.5. To prove (3), let $\Delta_{y,s}(t) = [f(y + su + tv) - f(y + tv)]/s$ so that we may write

$$\begin{aligned}
\limsup_{\substack{y \to x \\ s \to 0}} \frac{f'(y+su; v) - f'(y; v)}{s} &= \limsup_{\substack{y \to x \\ s \to 0}} \lim_{t \downarrow 0} \frac{1}{t} [\Delta_{y,s}(t) - \Delta_{y,s}(0)] \\
&\leqq \limsup_{\substack{y \to x \\ s,t \to 0}} \frac{1}{t} [\Delta_{y,s}(t) - \Delta_{y,s}(0)] \\
&= f^\infty(x; u, v).
\end{aligned}$$

Conversely, if $y \in X$ and $s, t \in \mathbb{R}_+$ we may use Lemma 1.4 to find $\alpha \in]0, t[$ so that with $y' = y + \alpha v$ we have

$$\frac{1}{t} [\Delta_{y,s}(t) - \Delta_{y,s}(0)] \leqq \frac{f'(y' + su; v) - f'(y'; v)}{s}.$$

Moreover, if $x = y + t_1 v$ and/or $x = y + su + t_2 v$, then we may always take $\alpha \in]0, t[$ different from t_1 and/or t_2 so that $y' \neq x$ and $y' + su \neq x$. We conclude by taking upper limits that

$$f^\infty(x; u, v) \leqq \limsup_{\substack{y \to x \\ s \downarrow 0}} \frac{f'(y+su; v) - f'(y; v)}{s},$$

with $y \neq x$ and $y + su \neq x$ in the upper limit. □

Next let us give some examples illustrating the notions introduced so far.

Example 1. It is easy to see that when f is linear we have $f^\infty(x; u, v) \equiv 0$ and $\partial^2 f(x)(u) = \{0\}$. Similarly, when $f(x) = b(x, x)$ with b a bilinear form, then $f^\infty(x; u, v) = b(u, v) + b(v, u)$ so that when b is continuous we have $\partial^2 f(x)(u) = \{b(\cdot, u) + b(u, \cdot)\}$.

Example 2. Let $f: X \to \mathbb{R}$ of class C^2 at x; then $f^\infty(x; u, v) = D^2 f(x)uv$ and $\partial^2 f(x)(u) = \{D^2 f(x)u\}$.

We recall that a fan A is said to be *linearly generated* if there exists a closed and convex family $\mathcal{A}$ of linear operators from X to X^* such that $A(u) = \{Lu: L \in \mathcal{A}\}$. With the above terminology we can say that the generalized Hessian of a C^2 function is linearly generated by the singleton $\{D^2 f(x)\}$. It is easy to see that any single-valued fan is linearly generated by one linear operator. It remains as an open question to

characterize the fans that are linearly generated and, in particular, for which class of functions the generalized Hessian enjoys this property. A partial answer will be given in § 2.

Example 3. If $f: X \to \mathbb{R}$ is sublinear and continuous, then

$$f^{\infty}(0; u, v) = \begin{cases} 0 & \text{if } v \in K(u), \\ +\infty & \text{otherwise,} \end{cases}$$

where $K(u)$ is the closed convex cone $\{v \in X: f'(x+u; v) \leqq f'(x; v)\ \forall x \in X\}$. To show this, observe that $f'(\alpha x; v) = f'(x; v)$ for all $x \in X$, $\alpha > 0$ and use formula (3). In this case the generalized Hessian turns out to be the polar set of $K(u)$, that is, $\partial^2 f(0)(u) = \{x^*: \langle x^*, v\rangle \leqq 0 \text{ for all } v \in K(u)\} = K(u)^0$.

An important particular case is $f(x) = \|x\|$ where $\|\cdot\|$ is the norm of a Hilbert space H. A direct calculation gives $K(u) = \{-tu: t \geqq 0\}$ and $\partial^2 f(0)(u) = \{v \in H: \langle v, u\rangle \geqq 0\}$. In particular, for $H = \mathbb{R}$ and $\|\cdot\| = |\cdot|$ we get $|\cdot|^{\infty}(0; u, v) = 0$ if $uv \leqq 0$ and $+\infty$ otherwise. Using this we may also conclude for $f(x_1, \cdots, x_n) = \sum_{i=1}^n |x_i|$ that

$$f^{\infty}(0; u, v) = \begin{cases} 0 & \text{if } u_i v_i \leqq 0 \text{ for } i = 1, \cdots, n, \\ +\infty & \text{otherwise,} \end{cases}$$

and $\partial^2 f(0)(u) = \{v: v_i u_i \geqq 0 \text{ for all } i = 1, \cdots, n\}$.

Example 4. Finally, let us show a few simple examples for $X = \mathbb{R}$. The first, $f(x) = \frac{1}{2}x|x|$, is a $C^{1,1}$ function that is not twice differentiable at zero. We have $f^{\infty}(0; u, v) = |uv|$ and $\partial^2 f(0)(u) = [-u, u]$.

The second, $f(x) = \sqrt{|x|+1}$ is not even differentiable at zero but using formula (3) it is easy to show that $f^{\infty}(0; u, v) = -\frac{1}{4}uv$ if $uv \leqq 0$ and $+\infty$ otherwise. This example shows moreover that (4) can fail if f is not continuously Gâteaux differentiable.

It is known [22] that there is a one-to-one correspondence between the family of all fans and the set of all bisublinear functions that are lower semicontinuous (l.s.c.) in the second variable. This correspondence is stated via the support *support function* of a fan.

Namely, to every fan $A: X \rightrightarrows X^*$ we associate the bisublinear function $S_A: X \times X \to \bar{\mathbb{R}}$ defined by $S_A(u, v) = \sup \langle A(u), v\rangle$. Conversely, to every bisublinear function $S: X \times X \to \bar{\mathbb{R}}$ that is l.s.c. on the second variable corresponds a unique fan $A: X \rightrightarrows X^*$ such that $S_A = S$, which may be characterized by $A(u) = \{x^* \in X^*: \langle x^*, v\rangle \leqq S(u, v) \text{ for all } v \in X\}$. This discussion motivates the following definition.

DEFINITION 1.6. A function $f: X \to \mathbb{R}$ is called *twice C-differentiable* at x if $f^{\infty}(x; u, \cdot)$ (or equivalently $f^{\infty}(x; \cdot, u)$) is l.s.c. for each $u \in X$.

In other words, f is twice C-differentiable at x if $\partial^2 f(x)$ is a fan and its support function is $f^{\infty}(x; \cdot, \cdot)$; that is,

$$f^{\infty}(x; u, v) = \sup \langle \partial^2 f(x)(u), v\rangle. \tag{9}$$

Note that all the examples presented so far are twice C-differentiable functions. This property fails, for instance, with $f(x) = x^{4/3}$ where $\partial^2 f(0)(u)$ is empty for all $u \neq 0$.

The differential concept most naturally linked to the notion of generalized Hessian is the one of twice strict differentiability.

DEFINITION 1.7. A function $f: X \to \mathbb{R}$ is called *twice strictly differentiable* at x if there exists a linear operator $D^2 f(x): X \to X^*$ such that

$$\langle D^2 f(x)(u), v\rangle = \lim_{\substack{y \to x \\ s,t \to 0}} \frac{f(y+su+tv) - f(y+su) - f(y+tv) + f(y)}{st}. \tag{10}$$

PROPOSITION 1.8. *A function $f: X \to \mathbb{R}$ is twice strictly differentiable at x if and only if f is twice C-differentiable at x with $\partial^2 f(x)$ single valued. In such a case we have $\partial^2 f(x)(u) = \{D^2 f(x)u\}$.*

Proof. The "only if" part is evident. Conversely, as pointed out in Example 2, when $\partial^2 f(x)$ is single valued, it is generated by a linear operator that we will denote $D^2 f(x)$. Let us prove that (10) holds.

From (9) we have

$$f^\infty(x; u, v) = \langle D^2 f(x)u, v\rangle = -\langle D^2 f(x)(-u), v\rangle = -f^\infty(x; -u, v),$$

so that, with Δ_f^2 as in the proof of Proposition 1.2, we may write

$$\limsup_{\substack{y\to x\\ s,t\to 0}} \Delta_f^2(y, s, t, u, v) = f^\infty(x; u, v) = -f^\infty(x; -u, v) = \liminf_{\substack{y\to x\\ s,t\to 0}} \Delta_f^2(y, s, t, u, v),$$

and the result follows. □

2. The twice locally Lipschitzian case. To obtain some continuity properties for the generalized Hessian we introduce two generalizations of the locally Lipschitzian property of a function (Lebourg [24, p. 126]).

DEFINITION 2.1. A function $f: X \to \mathbb{R}$ is called *twice locally Lipschitzian* (twice l.l.) at x, if for each $v \in X$ there exist neighborhoods V of x and U of zero such that $f^\infty(V; U, v)$ is bounded in $\mathbb{R}$. If this boundedness is uniform in v, that is, if there exist neighborhoods V of x and U of zero such that $f^\infty(V; U, U)$ is bounded in $\mathbb{R}$, then f is said to be *twice uniformly locally Lipschitzian* (twice u.l.l.) at x.

The following technical lemma will be useful when studying the continuity properties of the multifunctions $y \to \partial^2 f(y)(u)$ and $(y, u) \to \partial^2 f(y)(u)$.

LEMMA 2.2. *Let $\{p_i: X \to \mathbb{R}\}_{i\in I}$ be a family of sublinear functions, for which there exists a neighborhood U of $0 \in X$ such that $\bigcup \{p_i(U): i \in I\}$ is bounded in $\mathbb{R}$. Then $\{p_i\}_{i\in I}$ is uniformly equicontinuous.*

Proof. Without loss of generality we suppose that U is balanced. Let M be the least upper bound of $\bigcup \{p_i(U): i \in I\}$. If we take $\varepsilon > 0$ and $i \in I$, then for every h, k satisfying $h - k \in \varepsilon M^{-1} U$ we have that

$$p_i(h) - p_i(k) \leqq p_i(h-k) \leqq \varepsilon M^{-1} \sup \bigcup \{p_i(U): i \in I\} = \varepsilon.$$

By a symmetrical argument we conclude that for all $i \in I$,

$$|p_i(h) - p_i(k)| \leqq \varepsilon$$

whenever $k - h \in W := \varepsilon M^{-1} U$, which establishes the result. □

If $f: X \to \mathbb{R}$ is twice l.l. at x, using the above lemma for $p_y = f^\infty(y; u, \cdot)$ and $I = V$ where V is chosen as in Definition 2.1 we see that f is twice C-differentiable at every point of V.

Let Y, Z be two t.v.s. and $A: Y \rightrightarrows Z$ a point-to-set map. We recall that A is said to be *locally compact* at $y \in Y$ if there exists a neighborhood V of y such that $A(V) = \bigcup_{y'\in V} A(y')$ is relatively compact. A is said to be *closed* at y if for every generalized sequence $y_\alpha \to y$ and $z_\alpha \to z$ with $z_\alpha \in A(y_\alpha)$ we have $z \in A(y)$. Finally, if A is locally compact and closed at y, then we say that A is *upper semicontinuous* at y.

PROPOSITION 2.3. *Assume that $f: X \to \mathbb{R}$ is twice* l.l. *at x. Then, for each $u \in X$ the following hold*:

(a) *$y \to \partial^2 f(y)(u)$ is locally w^*-compact, and a fortiori* u.s.c., *at x. In particular, $\partial^2 f(x)(u)$ is w^*-compact.*

(b) *$(y, w) \to f^\infty(y; w, v)$ is* u.s.c. *at (x, u) for each $v \in X$, and the multifunction $(y, w) \to \partial^2 f(y)(w)$ is closed at (x, u).*

(c) *If in addition, f is twice* u.l.l. *at x, then* $(y, w) \to \partial^2 f(y)(w)$ *is locally* w^**-compact, and a fortiori* u.s.c., *at* (x, u).

Proof. (a) Let $u \in X$ and choose neighborhoods V of x and U of zero such that $f^\infty(V; U, u)$ is bounded, so that $\{f^\infty(y; u, \cdot): y \in V\}$ is uniformly equicontinuous by the previous lemma.

Now, if $x^* \in \partial^2 f(V)(u)$ then we have for some $y \in V$

$$-f^\infty(y; u, -v) \leqq \langle x^*, v\rangle \leqq f^\infty(y; u, v) \quad \text{for all } v \in X,$$

which proves that $\partial^2 f(V)(u)$ is an equicontinuous family of linear functionals, and henceforth is w^*-relatively compact. The upper semicontinuity follows immediately from Proposition 1.2(b).

(b) Let us choose arbitrary nets $x_\alpha \to x$ and $u_\alpha \to u$. The uniform equicontinuity of $\{f^\infty(y; \cdot, v): y \in V\}$, for a suitable neighborhood V of x, shows that

$$\lim_\alpha [f^\infty(x_\alpha; u_\alpha, v) - f^\infty(x_\alpha; u, v)] = 0,$$

from which we conclude, using Proposition 1.2(b), the inequality

$$\limsup_\alpha f^\infty(x_\alpha; u_\alpha, v) = \limsup_\alpha f^\infty(x_\alpha; u, v) \leqq f^\infty(x; u, v).$$

The closedness of $(y, w) \to \partial^2 f(y)(w)$ at (x, u) follows immediately by a similar argument to that given in Proposition 1.2(b).

(c) Let us consider neighborhoods V of x and U of zero such that $f^\infty(V; U, U)$ is bounded, so that $\{f^\infty(y; w, \cdot): y \in V, w \in U\}$ is uniformly equicontinuous by the previous lemma.

For $u \in X$, let $r > 0$ be such that $ru \in \operatorname{int}(U)$, so that $U' = U/r$ is a neighborhood of u. It is easy to see that $\{f^\infty(y; w, \cdot): y \in V, w \in U'\}$ is still uniformly equicontinuous. Now, by a similar argument to the one used in (a), we conclude the equicontinuity of $\partial^2 f(V)(U')$ and henceforth its w^*-relative compactness. □

An important class of twice uniformly locally Lipschitzian functions is the $C^{1,1}$, that is, Gâteaux differentiable functions with locally Lipschitzian gradient, that is, verifying that, for each $x \in X$ there exist a neighborhood V of x, a continuous seminorm p and a neighborhood U of 0 such that

$$|\langle \nabla f(y) - \nabla f(z), v\rangle| \leqq p(y - z) \quad \text{for all } v \in U \text{ and } y, z \in V.$$

In a recent work by Milosz [26], the same kind of generalized derivatives are considered, and it is shown that in the finite-dimensional case the finiteness of f^∞ is equivalent to f being of class $\mathscr{C}^{1,1}$. In a companion technical note [12], these results are extended to the case of normed spaces by showing that twice uniform local Lipschitzianity and $\mathscr{C}^{1,1}$ are equivalent properties.

Another approach to second-order generalized differentiation in the finite-dimensional setting was proposed by Hiriart-Urruty and developed by Araya and Gormaz [1] and Hiriart-Urruty, Strodiot, and Hien Nguyen [19]. The basic idea is to use a theorem by Rademacher that states that a locally Lipschitzian function between finite-dimensional spaces is differentiable Lebesgue almost everywhere, to define for a $C^{1,1}$ function $f: \mathbb{R}^n \to \mathbb{R}$ the generalized Hessian matrix as the compact nonempty convex set

$$\partial_H^2 f(x) = \operatorname{co}\{\lim \nabla^2 f(y_k): y_k (\in \operatorname{dom}(\nabla^2 f)) \to x\}, \tag{11}$$

where co denotes convex hull. It is easily shown that

$$\sup \langle \partial_H^2 f(x)u, v\rangle = \limsup_{\substack{y \to x \\ t \to 0}} t^{-1}\langle \nabla f(y + tu) - \nabla f(y), v\rangle,$$

and from (3) and (9) we conclude

$$\partial^2 f(x)(u) = \partial_H^2 f(x)u,$$

so that $\partial_H^2 f(x)$ generates the fan $\partial^2 f(x)$. This is a partial answer to the question posed in Example 2.

Closely related to Hiriart-Urruty's and our approach, is the notion of upper derivative in the sense of Ioffe [22, Def. 9.7]. In fact, for an arbitrary function $F: X \to Y$, where X and Y are Banach spaces, Ioffe defines the upper derivative of F at x as the fan $D^0F(x): Y^* \rightrightarrows X^*$ whose support function is

$$F^0(x; y^*, v) = F^0_{y^*}(x; v),$$

where $F_{y^*}(\cdot) = \langle y^*, F(\cdot)\rangle$ (this makes sense when $F^0(x; \cdot, v)$ is w^*-lower semicontinuous). Hence, if we take a differentiable function $f: X \to \mathbb{R}$, the upper derivative of $\nabla f: X \to X^*$ will be precisely what we have called the generalized Hessian of f, which by [22, Prop. 10.9] when $X = \mathbb{R}^n$ and $f \in C^{1,1}$ will be linearly generated by $\partial_H^2 f(x)$ as noted previously.

On the other hand, the symmetry of $f^\infty(x; \cdot, \cdot)$ implies that the generalized Hessian $\partial^2 f(x)$ also coincides with the C-coderivative of ∇f at x [22, Def. 9.12], which is the adjoint fan of $D^0(\nabla f)(x)$ and is denoted by $D^0_C(\nabla f)(x)$. Hence we have, for a differentiable function $f: X \to \mathbb{R}$ that is twice C-differentiable, that

$$\partial^2 f(x) = D^0(\nabla f)(x) = D^0_C(\nabla f)(x).$$

This implies in turn, by [22, Prop. 10.7], that $\partial^2 f(x)$ is linearly generated if $f: X \to \mathbb{R}$ is $C^{1,1}$, for a general Banach space X. This result allows the definition of the (linear) generalized Hessian as

$$\partial_L^2 f(x) = \{A: X \to X^*: A \text{ is linear and } Au \in \partial^2 f(x)(u),\ \forall u \in X\}.$$

Obviously enough, every result concerning $\partial^2 f(x)$ can be restated for $C^{1,1}$ functions in terms of $\partial_L^2 f(x)$.

3. Some calculus rules. To make $f^\infty(x; v, u)$ and $\partial^2 f(x)(u)$ computable, we now establish some calculus rules that permit the calculation of them when the function f is built up from "simpler" functions through sums, composition, and maximum.

DEFINITION 3.1. We say that $f: X \to \mathbb{R}$ is *twice subregular* at x if the limit

$$f''(x; u, v) = \lim_{\substack{s,t \to 0 \\ st > 0}} \frac{f(x+su+tv) - f(x+su) - f(x+tv) + f(x)}{st} \tag{12}$$

exists (eventually $\pm\infty$) and is equal to $f^\infty(x; u, v)$.

It is easy to verify that linear, quadratic, and strictly differentiable functions are examples of twice subregular functions. The same is true for the norm in a Hilbert space.

The above regularity property will be used in the following propositions to ensure equality in the formulas we present.

PROPOSITION 3.2. *Let $f: X \to \mathbb{R}$ and $r \in \mathbb{R}\backslash\{0\}$. Then we have for $x \in X$*

$$\partial^2(rf)(x) = r\partial^2 f(x), \tag{13}$$

where $r\partial^2 f(x)$ is defined as $[r\partial^2 f(x)](u) = r[\partial^2 f(x)(u)]$.

Proof. From Proposition 1.2(a) and (c) it follows directly that

$$(rf)^\infty(x; u, v) = |r|(sg(r)f)^\infty(x; u, v) = f^\infty(x; u, |r|sg(r)v) = f^\infty(x; u, rv),$$

so that

$$\begin{aligned}\partial^2(rf)(x)(u) &= \{x^*: \langle x^*, v\rangle \leqq f^\infty(x; u, rv),\ \forall v \in X\}\\ &= \{x^*: \langle x^*/r, v\rangle \leqq f^\infty(x; u, v),\ \forall v \in X\}\\ &= r\partial^2 f(x)(u).\end{aligned}$$ □

The next proposition deals with the generalized Hessian of a sum of two functions. Previously, let us recall that the sum of two prefans A and B is defined as the prefan $\overline{(A+B)}(u) = \overline{A(u)+B(u)}$.

PROPOSITION 3.3. *Let $f, g: X \to \mathbb{R}$ and $x \in X$; then we have*

$$(f+g)^\infty(x; u, v) \leqq f^\infty(x; u, v) + g^\infty(x; u, v), \tag{14}$$

and if f and g are twice C-differentiable at x we get

$$\partial^2(f+g)(x) \subset \overline{\partial^2 f(x) + \partial^2 g(x)}. \tag{15}$$

Equality holds in (14) *and* (15) *when in addition f or g is twice strictly differentiable at x, and also if both f and g are twice subregular at x. In the last case we also have that $f+g$ is twice subregular at x.*

Proof. Inequality (14) is obvious from the definition of $(f+g)^\infty$ and the subadditivity of the upper limit. The same can be said for the equality when f or g are strictly differentiable at x.

Now, when f and g are twice C-differentiable, $f^\infty(x; u, \cdot) + g^\infty(x; u, \cdot)$ is the support functional of the closed convex set $\partial^2 f(x)(u) + \partial^2 g(x)(u)$. Thus, the support functional of $\partial^2(f+g)(x)$ is bounded above by $(f+g)^\infty(x; u, v)$ and a fortiori by the support functional of $\overline{\partial^2 f(x)(u) + \partial^2 g(x)(u)}$, so that a classical result of convex analysis permits us to conclude (15).

The same argument can be used to derive the equality in (15) when (14) holds with equality by noting that in this case $(f+g)$ is twice C-differentiable at x whenever f and g are, and henceforth the support functional of the set $\partial^2(f+g)(x)(u)$ is exactly $(f+g)^\infty(x; u, \cdot)$.

It remains to verify that under the stronger assumption that f and g are twice subregular, equality holds in (14). Indeed, in such a case we have

$$\begin{aligned}(f+g)^\infty(x; u, v) &\leqq f^\infty(x; u, v) + g^\infty(x; u, v)\\ &= f''(x; u, v) + g''(x; u, v)\\ &= (f+g)''(x; u, v),\end{aligned}$$

but since always $(f+g)''(x; u, v) \leqq (f+g)^\infty(x; u, v)$, the result follows. □

Next let us turn to the problem of calculating the generalized Hessian for the composition of two functions $f: X \to \mathbb{R}$ and $g: Y \to X$ where Y and X are l.c.t.v.s. paired with Y^* and X^*, respectively.

PROPOSITION 3.4. *Let the above f be of class $C^{1,1}$ at $x = g(y)$ and g twice continuously differentiable at y. Then the following chain rule holds:*

$$\partial^2(f \circ g)(y)(u) \subset Dg(y)^* \partial^2 f(g(y))(Dg(y)u) + \nabla f(g(y)) \circ D^2 g(y)(u), \tag{16}$$

where Dg and $D^2 g$ are the first and second Gâteaux derivatives of g, and * denotes the adjoint map of the corresponding linear operator. Equality holds in (16) *if g is open at y.*

In particular, if g is the linear operator $A: Y \to X$ then formula (16) *becomes*

$$\partial^2(f \circ A)(y)(u) \subset A^* \partial^2 f(Ay)(Au), \tag{17}$$

with equality if A is open.

Proof. To prove (16) it is sufficient to show that the support functionals of the sets in that formula satisfy the inequality

$$(f\circ g)^{\infty}(y;u,v)\leqq f^{\infty}(g(y);Dg(y)u,Dg(y)v)+\langle\nabla f(g(y)),D^2g(y)(u)v\rangle,$$

for each $v\in Y$. To this end, let us compute $(f\circ g)^{\infty}$, by using characterization (3), that is to say

$$(f\circ g)^{\infty}(y;u,v)=\limsup_{\substack{z\to y\\ t\to 0}}\frac{1}{t}[\langle\nabla f(g(z+tu)),Dg(z+tu)v\rangle-\langle\nabla f(g(z)),Dg(z)v\rangle].$$

By adding and subtracting $\langle\nabla f(g(z)),Dg(z+tu)v\rangle$, and noting that being f of class $C^{1,1}$, we have

$$\lim_{\substack{z\to y\\ t\to 0}}\frac{1}{t}\langle\nabla f(g(z)),Dg(z+tu)v-Dg(z)v\rangle=\langle\nabla f(g(y)),D^2g(y)(u)v\rangle,$$

we conclude

$$(f\circ g)^{\infty}(y;u,v)=\limsup_{\substack{z\to y\\ t\to 0}}\frac{1}{t}\langle\nabla f(g(z+tu))-\nabla f(g(z)),Dg(z+tu)v\rangle$$
$$+\langle\nabla f(g(y)),D^2g(y)(u)v\rangle.$$

The proposition will be proved if we show that the previous upper limit is bounded above by (and if g is open at y, equal to) $f^{\infty}(g(y);Dg(y)u,Dg(y)v)$.

To do this let us observe again that by adding and subtracting $\langle\nabla f(g(z)+tDg(y)u),Dg(z+tu)v\rangle$, and since

$$\lim_{\substack{z\to y\\ t\to 0}}\frac{1}{t}[g(z+tu)-g(z)-tDg(y)u]=0,$$

we conclude, by using the fact that f is $C^{1,1}$, that

$$\lim_{\substack{z\to y\\ t\to 0}}\frac{1}{t}\langle\nabla f(g(z+tu))-\nabla f(g(z)+tDg(y)u),Dg(z+tu)v\rangle=0,$$

and then

$$(f\circ g)^{\infty}(y;u,v)=\limsup_{\substack{z\to y\\ t\to 0}}\frac{1}{t}\langle\nabla f(g(z)+tDg(y)u)-\nabla f(g(z)),Dg(z+tu)v\rangle \tag{18}$$
$$+\langle\nabla f(g(y)),D^2g(y)(u)v\rangle.$$

To proceed with the proof we need the following lemma.

LEMMA 3.5. *Let* $x_\alpha(\in X)\to x$, $t_\alpha(\in\mathbb{R})\to 0$, *and* $w_\alpha(\in X)\to 0$; *then*

$$\lim_{\alpha}\frac{1}{t_\alpha}\langle\nabla f(x_\alpha+t_\alpha u)-\nabla f(x_\alpha),w_\alpha\rangle=0.$$

Proof. Since f is $C^{1,1}$ at x, let us choose p a continuous seminorm, V a neighborhood of x, and U a neighborhood of zero such that

$$|\langle\nabla f(y)-\nabla f(z),w\rangle|\leqq p(y-z)\quad\text{for all } y,z\in V,\ w\in U.$$

Defining $r_\alpha=\sup\{r: rw_\alpha\in U\}$, it is easy to see that $r_\alpha\to+\infty$ (since $w_\alpha\to 0$).

Then, for sufficiently large α we have

$$\left|\frac{1}{t_\alpha}\langle \nabla f(x_\alpha + t_\alpha u) - \nabla f(x_\alpha), r_\alpha w_\alpha\rangle\right| \leqq \left|\frac{1}{t_\alpha}\right| p(t_\alpha u) = p(u),$$

so that dividing this inequality by r_α and passing to the limit we obtain the desired conclusion. □

Using this lemma we may continue our proof by noting that the term $Dg(z+tu)v$ in (18) can be replaced by $Dg(y)u$, so that

$$(f\circ g)^\infty(y; u, v) = \limsup_{\substack{z\to y\\ t\to 0}} \frac{1}{t}\langle \nabla f(g(z)+tDg(y)u) - \nabla f(g(z)), Dg(y)v\rangle$$
$$+\langle \nabla f(g(y)), D^2 g(y)(u)v\rangle.$$

Now, $g(z)\to g(y)$ when $z\to y$, so that we have

$$(f\circ g)^\infty(y; u, v) \leqq \limsup_{\substack{z\to g(y)\\ t\to 0}} \frac{1}{t}\langle \nabla f(z+tDg(y)u) - \nabla f(z), Dg(y)v\rangle$$
$$+\langle \nabla f(g(y)), D^2 g(y)(u)v\rangle,$$

with equality if g maps the neighborhoods of y into the neighborhoods of $g(y)$.

To conclude we may just use characterization (3) to obtain

$$(f\circ g)^\infty(y; u, v) \leqq f^\infty(g(y); Dg(y)u, Dg(y)v) + \langle \nabla f(g(y)), D^2 g(y)(u)v\rangle,$$

with equality if g is open at y. □

Other chain rules have been developed for $C^{1,1}$ functions in finite-dimensional spaces in the paper by Hiriart-Urruty, Strodiot, and Hien Nguyen [19]. Let us mention that any effort made to weaken the hypothesis made on f in the previous result would be worthy.

Another important question is to derive calculus rules for functions of the max type. Namely, let us suppose we are given a finite family of C^2 functions $f_i: X\to\mathbb{R}$ for $i\in I=\{1,\cdots,n\}$ and consider the mapping $f: X\to\mathbb{R}$ given by

$$f(x) = \max_{i\in I} f_i(x).$$

It is well known that f is directionally differentiable with

$$f'(x; v) = \max_{i\in I(x)} f_i'(x; v),$$

where $I(x)=\{i\in I: f_i(x)=f(x)\}$. Let us set

$$I(x, v) = \{i\in I(x): f_i'(x, v) = f'(x, v)\}.$$

DEFINITION 3.6. With the above notation, we will say that an index $i\in I$ is *essential* at x if there exists a net $x_\alpha\to x$ with $I(x_\alpha)=\{i\}$. We will denote by $I^*(x)$ the set of essential indexes at x.

It is clear that $I^*(x)\subset I(x)$ and that for a local representation of f at x it suffices to consider the functions $\{f_i: i\in I^*(x)\}$. Now, the computation of $I^*(x)$ may not be an easy task, but we can mention the following practical criteria.

PROPOSITION 3.7. *If $i\in I(x)$ is such that there exists $v\in X$ with $I(x, v)=\{i\}$, then $i\in I^*(x)$. In particular, if $\{\nabla f_j(x): j\in I(x)\}$ is linearly independent, or more generally, affinely independent, then $I^*(x)=I(x)$.*

Proof. $I(x, v)=\{i\}$ means $f'(x; v)=f_i'(x; v)>f_j'(x; v)$ for all $j\in I(x)\backslash\{i\}$. Therefore, $f(x+tv)=f_i(x+tv)>f_j(x+tv)$ for all t sufficiently small and consequently $I(x+tv)=\{i\}$.

Now, if $\{\nabla f_j(x): j\in I(x)\}$ is affinely independent, then for each $i\in I(x)$ we may find $v\in X$ such that

$$\langle\nabla f_i(x)-\nabla f_j(x), v\rangle=+1 \quad \forall j\in I(x)\backslash\{i\},$$

which implies $I(x, v)=\{i\}$. Since $I^*(x)\subset I(x)$ is always satisfied, the proof is complete. □

We may now prove the following easy result.

PROPOSITION 3.8. *With the above notation we have for each* $(u, v)\in X\times X$

$$\max_{i\in I^*(x)} D^2f_i(x)uv\leqq f^\infty(x; u, v), \tag{19}$$

and

$$\operatorname{co}\{D^2f_i(x)u: i\in I^*(x)\}\subset\partial^2 f(x)(u), \tag{20}$$

so that $\mathscr{A}=\operatorname{co}\{D^2f_i(x): i\in I^*(x)\}$ *is a set of linear selections of* $\partial^2 f(x)$.

Proof. Clearly, it suffices to show (19). Now, take $i\in I^*(x)$ and select a net $x_\alpha\to x$ with $I(x_\alpha)=\{i\}$. Then, near x_α we have $f\equiv f_i$ and therefore

$$D^2f_i(x_\alpha)uv=f^\infty(x_\alpha; u, v).$$

By going to the limit and using the upper semicontinuity of $f^\infty(\cdot; u, v)$ we obtain

$$D^2f_i(x)uv\leqq f^\infty(x; u, v),$$

and (19) follows at once. □

Equality in (19) is hopeless in general as we often have $f^\infty(x; u, v)=\infty$, as the following proposition shows.

PROPOSITION 3.9. *Suppose* $\{\nabla f_i(x)\}_{i\in I(x)}$ *is affinely independent, and consider the following condition on* (u, v):

$$H(u, v): \langle\nabla f_i(x)-\nabla f_j(x), v\rangle\langle\nabla f_i(x)-\nabla f_j(x), u\rangle\leqq 0 \quad \textit{for all } i, j\in I(x).$$

Then $f^\infty(x; u, v)=+\infty$ *whenever* $H(u, v)$ *does not hold.*

Proof. Let us take $i, j\in I(x)$, violating the inequality in $H(u, v)$.

Since $f^\infty(x; -u, -v)=f^\infty(x; u, v)$ we may assume (by eventually changing u to $-u$ and v to $-v$) that

$$\langle\nabla f_i(x), v\rangle>\langle\nabla f_j(x), v\rangle \quad \text{and} \quad \langle\nabla f_i(x), u\rangle>\langle\nabla f_j(x), u\rangle.$$

Now, select $h\in X$ with

$$\langle\nabla f_i(x), h\rangle=\langle\nabla f_j(x), h\rangle=\langle\nabla f_k(x), h\rangle+1 \quad \text{for all } k\in I(x)\backslash\{i, j\}.$$

Using a standard implicit function theorem we may find a path $x(t)\in X$ such that $x(0)=x$, $x'(0)=h$ and verifying $f_i(x(t))=f_j(x(t))>f_k(x(t))$ for every $k\in I(x)\backslash\{i, j\}$ and $t>0$ small enough, that is, $I(x(t))=\{i, j\}$.

Since $\langle\nabla f_i(x), u\rangle>\langle\nabla f_j(x), u\rangle$, the same holds with x replaced by $x(t)$ provided $t>0$ is small, and we may then find $\varepsilon_t>0$ such that

$$I\left(x(t)+\frac{s}{2}u\right)=\{i\} \quad \text{and} \quad I\left(x(t)-\frac{s}{2}u\right)=\{j\}$$

for all $s\in\,]0, \varepsilon_t[$ and t near zero.

Take $s(t) \in]0, \varepsilon_t[$ with $\lim_{t\downarrow 0} s(t) = 0$ and set $y(t) = x(t) - (s(t)/2)u$. Then

$$\lim_{t\downarrow 0} f'(y(t)+s(t)u;\, v) = \lim_{t\downarrow 0} \langle \nabla f_i(y(t)+s(t)u),\, v\rangle = \langle \nabla f_i(x),\, v\rangle,$$

$$\lim_{t\downarrow 0} f'(y(t);\, v) = \lim_{t\downarrow 0} \langle \nabla f_j(y(t)),\, v\rangle = \langle \nabla f_j(x),\, v\rangle,$$

and since $\langle \nabla f_i(x), v\rangle > \langle \nabla f_j(x), v\rangle$ we conclude from Proposition 1.3 that

$$f^\infty(x;\, u,\, v) \geqq \limsup_{t\downarrow 0} \frac{f'(y(t)+s(t)u;\, v) - f'(y(t);\, v)}{s(t)} = +\infty. \qquad \square$$

In view of the previous results, a natural conjecture would be that the following characterization holds:

$$\text{(C)} \qquad f^\infty(x;\, u,\, v) = \begin{cases} \max\limits_{i\in I(x)} D^2 f_i(x)(u)v & \text{if } H(u,\, v) \text{ holds,} \\ +\infty & \text{otherwise,} \end{cases}$$

under the affine independence of $\{\nabla f_i(x)\}_{i\in I(x)}$. However, we have just been able to prove this when the f_i's are linear. (Note that by an appropriate translation of the origin this covers the affine case.)

PROPOSITION 3.10. *Let $(l_i)_{i\in I}$ be a finite set of affinely independent linear functions. Then, for $f(x) = \max_{i\in I} l_i(x)$ we have*

$$f^\infty(0;\, u,\, v) = \begin{cases} 0 & \text{if } (l_i - l_j)(v)(l_i - l_j)(u) \leqq 0 \quad \text{for all } i, j \in I, \\ +\infty & \text{otherwise.} \end{cases}$$

Proof. We must only prove that when $(l_i - l_j)(v)(l_i - l_j)(u) \leqq 0$ for all $i, j \in I$ then $f^\infty(0;\, u,\, v) = 0$. Now, from Example 3 we know that

$$f^\infty(0;\, u,\, v) = \begin{cases} 0 & \text{if } f'(x+u;\, v) \leqq f'(x;\, v) \quad \text{for all } x \in X, \\ +\infty & \text{otherwise,} \end{cases}$$

so we only need to prove $f'(x+u;\, v) \leqq f'(x;\, v)$ for all $x \in X$, that is,

$$l_i(v) \leqq \max\{l_j(v) : j \in I(x)\} \quad \text{for each } i \in I(x+u).$$

This is clear if $i \in I(x)$. Otherwise, for each $j \in I(x)$ we have $l_i(x) < l_j(x)$ and since $l_i(x+u) \geqq l_j(x+u)$ we deduce $l_i(u) > l_j(u)$. Consequently, $l_i(v) \leqq l_j(v)$ from our hypothesis, which shows in fact that

$$l_i(v) \leqq \min\{l_j(v) : j \in I(x)\} \quad \text{for all } i \in I(x+u),$$

and the proof is complete. $\square$

Example 5. Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x_1, \cdots, x_n) = \max\{x_1, \cdots, x_n\}$. Then we have

$$f^\infty(0;\, u,\, v) = \begin{cases} 0 & \text{if } (u_i - u_j)(v_i - v_j) \leqq 0 \quad \text{for } 1 \leqq i, j \leqq n, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that our conjecture (C) could be proved using this example if we had a chain rule stronger than Proposition 3.4 (for f directionally differentiable and twice-C-differentiable for instance).

4. Second-order Taylor expansion.

PROPOSITION 4.1. *Let $f : X \to \mathbb{R}$ be continuously Gâteaux differentiable and twice C-differentiable at every point of the segment $[x, y] \subset X$. Then there exists $\xi \in]x, y[$ such that*

$$f(y) \in f(x) + \langle \nabla f(x),\, y - x\rangle + \tfrac{1}{2}\overline{\langle \partial^2 f(\xi)(y-x),\, (y-x)\rangle}, \tag{21}$$

the closure being superfluous if f is $C^{1,1}$ in $[x, y]$.

Proof. Let us consider $h:\mathbb{R}\to\mathbb{R}$ defined as

$$h(t)=f(y+t(x-y))+t\langle\nabla f(y+t(x-y)),y-x\rangle+\tfrac{1}{2}at^2-f(y), \tag{22}$$

where a has been chosen so that $h(0)=h(1)=0$.

From Lemma 1.4, there exists $\alpha\in]0,1[$ such that $h^0(\alpha;s)\geqq 0$ for each $s\in\mathbb{R}$. Letting $h_1(t)=t\langle\nabla f(y+t(x-y)),y-x\rangle$ and $\xi=y+\alpha(x-y)$, we get

$$0\leqq h^0(\alpha;s)=a\alpha s+s\langle\nabla f(\xi),x-y\rangle+h_1^0(\alpha;s).$$

Moreover, setting $h=x-y$ we may write

$$h_1^0(\alpha;s)=\limsup_{\substack{t\to\alpha\\ r\to 0}}\frac{t}{r}\langle\nabla f(y+th+rsh)-\nabla f(y+th),-h\rangle-s\langle\nabla f(\xi),h\rangle$$

$$\leqq\alpha f^\infty(\xi;s(x-y),y-x)-s\langle\nabla f(\xi),x-y\rangle,$$

and we deduce that

$$0\leqq a\alpha s+\alpha f^\infty(\xi;s(x-y),y-x)\quad\text{for all } s\in\mathbb{R},$$

which, used for $s=1$ and $s=-1$, gives us

$$-f^\infty(\xi;y-x,-(y-x))\leqq a\leqq f^\infty(\xi;y-x,y-x).$$

Hence $a\in\overline{\langle\partial^2 f(\xi)(y-x),y-x\rangle}$ and from (22) with $t=1$ we get formula (21). Finally, if f is $C^{1,1}$ at ξ, then $\partial^2 f(\xi)(y-x)$ is w^*-compact so that adherence in (21) is superfluous. □

At this time we do not know when formula (21) holds for nonsmooth functions by replacing $\nabla f(x)$ with the generalized gradient $\partial f(x)$.

The above result will be used in the next section to obtain second-order necessary and sufficient optimality conditions. Meanwhile, let us use it to study the relationship between $\partial^2 f$ and the convexity of f.

DEFINITION 4.2. A prefan $A:X\rightrightarrows X^*$ will be said to be *positively defined* (p.d.) (respectively, *weakly positively defined* (w.p.d.)) if its support functional S satisfies $-S(u,-u)\geqq 0$ (respectively, $S(u,u)\geqq 0$) for each $u\in X$. If the above inequality is strict for $u\neq 0$, A will be said to be strictly p.d. (respectively, strictly w.p.d.).

PROPOSITION 4.3. *Let $f:X\to\mathbb{R}$ be convex and twice C-differentiable; then for each $x\in X$ the generalized Hessian $\partial^2 f(x)$ is* p.d. *The converse holds whenever f is moreover continuously Gâteaux differentiable.*

Proof. As pointed out in the remark following Definition 1.1,

$$f^\infty(x;u,v)=\limsup_{\substack{y\to x\\ s,t\to 0^+}}\frac{f(y+su+tv)-f(y+su)-f(y+tv)+f(y)}{st},$$

so that the first part will be proved if we show that for each $y\in X$ and $x,t\in\mathbb{R}_+$ we have

$$f(y-su+tu)-f(y-su)-f(y+tu)+f(y)\leqq 0,$$

which is an easy consequence of the convexity inequalities

$$f(y)\leqq\frac{t}{s+t}f(y-su)+\frac{s}{s+t}f(y+tu),$$

$$f(y-su+tu)\leqq\frac{s}{s+t}f(y-su)+\frac{t}{s+t}f(y+tu).$$

For the converse, let us take $x, y \in X$ and use Proposition 4.1, to obtain

$$f(y)-f(x)-\langle \nabla f(x), y-x\rangle \in \tfrac{1}{2}\overline{\langle \partial^2 f(\xi)(y-x), y-x\rangle} \subset \mathbb{R}_+,$$

so that $f(y) \geqq f(x)+\langle \nabla f(x), y-x\rangle$ and f is convex since x and y were chosen arbitrarily. □

5. Second-order optimality conditions. Given a function $f: X \to \mathbb{R}$ we can formulate the following optimality conditions for the unconstrained minimization problem:

$$\text{(P)} \qquad \text{Minimize } \{f(x): x \in X\}.$$

PROPOSITION 5.1. *A necessary condition for $x \in X$ to be a solution of problem* (P) *is that $f^\infty(x; u, u) \geqq 0$ for each $u \in X$. If the function f is twice C-differentiable at x, this condition corresponds to the fact that $\partial^2 f(x)$ is* w.p.d.

Proof. For t sufficiently small we have $f(x+tu) \geqq f(x)$ and henceforth

$$f^\infty(x; u, u) \geqq \limsup_{t \to 0} \frac{f(x+tu)-f(x)-f(x)+f(x-tu)}{t^2} \geqq 0. \qquad \square$$

Example 6. The function $|x|^{3/2}$ has a minimum at $x=0$ and therefore satisfies the above second-order optimality condition (s.o.o.c.).

On the other hand, $-|x|^{3/2}$ satisfies the first-order optimality condition $f'(0)=0$. Nevertheless, it fails to satisfy the above s.o.o.c. at zero and can be rejected as a candidate for a minimum (it is, in fact, a maximum).

PROPOSITION 5.2. *Assume that $f: \mathbb{R}^n \to \mathbb{R}$ is $C^{1,1}$ and $x \in X$ satisfies the first-order optimality condition $\nabla f(x)=0$. Then a sufficient condition for x to be a strict local minimum of f is that $\partial^2 f(x)$ is strictly* p.d.

Proof. By using a classical argument we see that if it is not so, we may find a sequence $x_n \to x$ such that $x_n \neq x$ and $f(x_n) \leqq f(x)$ for all $n \in \mathbb{N}$, and with no loss of generality we may assume that for some $u \in X$

$$u_n = (x_n - x)/\|x_n - x\| \to u.$$

Now, using Proposition 4.1 we may conclude the existence of a sequence $\xi_n \in]x, x_n[$, such that

$$a_n = 2(f(x_n)-f(x))/\|x_n - x\|^2 \in \langle \partial^2 f(\xi_n)(u_n), u_n\rangle,$$

so we have found sequences $\xi_n \to x$, $u_n \to u$ and $x_n^* \in \partial^2 f(\xi_n)(u_n)$ such that $a_n = \langle x_n^*, u_n\rangle \leqq 0$.

By Proposition 2.3(c), we may assume (eventually extracting subsequences) that (x_n^*) converges to some $x^* \in \partial^2 f(x)(u)$. But this would imply that $\langle x^*, u\rangle = \lim \langle x_n^*, u_n\rangle \leqq 0$ and a fortiori that $-f^\infty(x; u, -u) \leqq 0$ contradicting the fact that $\partial^2 f(x)$ is strictly p.d. □

Similar results can be proved for the constrained problem

$$\text{(R)} \qquad \text{Minimize } \{f_0(x): x \in Q\},$$

where $Q = \{x \in X: f_i(x) \leqq 0 \text{ for all } i \in I = \{1, \cdots, n\}\}$ and the f_i's are continuous and Gâteaux differentiable for each $i = 0, 1, \cdots, n$.

First of all, let us recall [5] the definition of the cones

$$D(Q, x) = \{u \in X: \exists r > 0 \text{ such that } x +]0, r[u \subset Q\}$$

and

$$T(Q, x) = \{u \in X: \langle \nabla f_i(x), u\rangle \leqq 0, \forall i \in I(x)\},$$

where $I(x) = \{i \in I: f_i(x) = 0\}$.

Let us recall also that $x \in Q$ is called a *regular point* for the problem (R) if $T(Q, x) = \overline{D(Q, x)}$, and that when x is a regular point that is a minimum for problem (R), then the Kuhn-Tucker Theorem [5, Thm. 3.4] ensures the existence of $\lambda = (\lambda_1, \cdots, \lambda_n) \in \mathbb{R}^n_+$ such that

$$\nabla L_\lambda(x) = 0 \quad \text{and} \quad L_\lambda(x) = f_0(x), \tag{23}$$

where $L_\lambda(x) = f_0(x) + \sum_{i \in I} \lambda_i f_i(x)$. Every λ verifying (23) is called a *multiplier.*

To get s.o.o.c. for problem (R) we associate to each multiplier the set

$$Q(\lambda) = \{x \in Q: f_i(x) = 0 \text{ if } \lambda_i > 0 \text{ and } f_i(x) \leqq 0 \text{ if } \lambda_i = 0\}.$$

PROPOSITION 5.3. *Suppose $x \in Q$ is a regular point that is a minimum for problem* (R). *Then for each multiplier $\lambda \in \mathbb{R}^n_+$ and for each $u \in D(Q(\lambda), x)$ we have $L^\infty_\lambda(x; u, u) \geqq 0$.*

Proof. Clearly, we have

$$L^\infty_\lambda(x; u, u) \geqq \limsup_{t \to 0^+} \frac{1}{t} \langle \nabla L_\lambda(x + tu), u\rangle.$$

Now, if $u \in D(Q(\lambda), x)$ there exists $r > 0$ such that for $t \in]0, r[$ we have $x + tu \in Q(\lambda)$ and hence $L_\lambda(x + tu) = f_0(x + tu)$. Then, being x a minimum we have for t sufficiently small

$$L_\lambda(x + tu) = f_0(x + tu) \geqq f_0(x) = L_\lambda(x),$$

so that using the mean value theorem we get that for each $t > 0$ sufficiently small there exists $t' \in]0, t[$ such that

$$t' \langle \nabla L_\lambda(x + t'u) u\rangle \geqq 0,$$

from which it follows easily that $\limsup_{t \to 0^+} 1/t \langle \nabla L_\lambda(x + tu), u\rangle \geqq 0$ and henceforth $L^\infty_\lambda(x; u, u) \geqq 0$. □

Conversely, we can establish a sufficient s.o.o.c. in the finite-dimensional case. Let us define the Bouligand tangent cone to Q at $x \in Q$ as

$$B(Q, x) = \{u \in X: \exists t_\alpha \downarrow 0, u_\alpha \to u \text{ such that } x + t_\alpha u_\alpha \in Q\}.$$

It is easy to verify the inclusion $\overline{D(Q, x)} \subset B(Q, x)$ and when Q is defined by differentiable constraints that $B(Q, x) \subset T(Q, x)$.

PROPOSITION 5.4. *Let $f_i: \mathbb{R}^n \to \mathbb{R}$, $i = 0, 1, \cdots, n$ be $C^{1,1}$ functions. Then a sufficient condition for a point $x \in Q$ to be a strict local minimizer for problem* (R) *is that there exists a multiplier $\lambda \in \mathbb{R}^n_+$ such that $-L^\infty_\lambda(x; u, -u) > 0$ for each $u \in B(Q, x) \backslash \{0\}$.*

Proof. The proof is analogous to that of Proposition 5.2. □

6. Some ideas concerning Newton's method. In this section we point out some simple facts concerning the following natural extension of Newton's method for the minimization of a $C^{1,1}$ function $f: \mathbb{R}^n \to \mathbb{R}$.

(A) Starting from an arbitrary point $x_0 \in \mathbb{R}^n$ generate the sequences $(x_k)_{k\in\mathbb{N}}$ and $(h_k)_{k\in\mathbb{N}}$ by

(i) $h_k \in \mathbb{R}^n$ is any solution of the inclusion $0 \in \nabla f(x_k) + \partial^2 f(x_k)(h_k)$,

(ii) $x_{k+1} = x_k + h_k$.

PROPOSITION 6.1. *Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by algorithm* (A) *and let us assume that it is convergent to some $x \in \mathbb{R}^n$. Then $\nabla f(x) = 0$ and furthermore, for k sufficiently large we have*

$$\|\nabla f(x_k)\| \leqq l\|x_{k+1} - x_k\| \tag{24}$$

for some constant $l \in \mathbb{R}_+$.

Proof. Since f is of class $\mathscr{C}^{1,1}$ at x, it is easy to see that for some $l \in \mathbb{R}_+$ and all y sufficiently close to x we have

$$\|x^*\| \leqq l\|h\| \quad \text{for all } x^* \in \partial^2 f(y)(h),$$

so that (24) follows at once by taking $x^* = -\nabla f(x_k)$ and $h = x_{k+1} - x_k$.

Then, the continuity of ∇f implies that $\nabla f(x) = 0$ and the proof is complete. □

It is clear, from what is known about Newton's method in the smooth case, that at most we can expect local convergence for algorithm (A) and nothing can be said about the cluster points of the sequence (x_k) (assuming there are any).

On the other hand, it remains as an open question to state *rate of convergence and local convergence* results for algorithm (A). In this sense, it is interesting to investigate the modification of algorithm (A), which consists of replacing (ii) $x_{k+1} = x_k + h_k$ by (ii)′ $x_{k+1} = x_k + \alpha_k h_k$ where α_k is chosen by directional exact or approximate minimization.

To conclude this section let us say a few words in connection with the solvability of the inclusion

$$\text{Find } h \in \mathbb{R}^n \text{ such that } 0 \in \nabla f(x) + \partial^2 f(x)(h), \tag{25}$$

which must be solved at each iteration.

The classical assumption in Newton's method is that the Hessian Matrix $\nabla^2 f(x)$ of f at x is nonsingular, which in our case could be stated as "There exists a nonsingular linear selection $A: \mathbb{R}^n \to \mathbb{R}^n$ of the generalized Hessian $\partial^2 f(x)$ (i.e., such that $A(h) \in \partial^2 f(x)(h)$ for all $h \in \mathbb{R}^n$)."

Nevertheless, the multivalued character of the generalized Hessian $\partial^2 f(x)$ and the nice properties enjoyed by $\partial^2 f(x)(\cdot)$, suggest that (25) is likely to have a solution (or even many of them).

The following result gives a sufficient condition for (25) to have at least one solution.

PROPOSITION 6.2. *Let f be of class $\mathscr{C}^{1,1}$ and suppose that for some $r > 0$ we have*

$$\|v\| = r \Rightarrow \langle \nabla f(x), v\rangle + f^\infty(x; v, v) \geqq 0. \tag{26}$$

Then there exists $h \in B(0, r)$ such that $0 \in \nabla f(x) + \partial^2 f(x)(h)$.

Proof. This is a consequence of Corollary 3 in [4, Chap. 6, § 4]. □

It is easy to see that when $\partial^2 f(x)$ is strictly w.p.d., condition (26) will be automatically satisfied for all $r \geqq \|\nabla f(x)\|/\alpha$, where

$$\alpha = \min_{\|v\|=1} f^\infty(x; v, v) > 0.$$

A final remark in connection with inclusion (25) is that it is equivalent to find $h \in \mathbb{R}^n$ such that $\langle \nabla f(x), v\rangle + f^\infty(x; h, v) \geqq 0$ for all $v \in \mathbb{R}^n$, so that any solution $h \in \mathbb{R}^n$ will satisfy, in particular,

$$\langle \nabla f(x), h\rangle \leqq f^\infty(x; h, -h),$$

and, in turn, whenever $\partial^2 f(x)$ is strictly p.d., we will have

$$\langle \nabla f(x), h\rangle < 0,$$

showing that h is a descent direction.

Example 7. An interesting case of minimization of a $\mathscr{C}^{1,1}$ function is provided by the so-called multiplier methods for solving

$$\text{Minimize } \{f_0(x): x \in \mathbb{R}^n, f_i(x) \leqq 0;\ i = 1, \cdots, p\},$$

which perform a sequence of unconstrained minimizations (with respect to x) of the augmented Lagrangian

$$L_c(x, y) = f_0(x) + \frac{1}{2c} \sum_{i=1}^{n} [y_i + cf_i(x)]_+^2 - y_i^2$$

where c and y are updated from one iteration to the next according to different specific rules.

It is easy to see that when the f_i's are $\mathscr{C}^2$ then L_c, and therefore $L_c(\cdot, y)$, is of class $\mathscr{C}^{1,1}$. Moreover, letting

$$H_c^i(x, y) = [y_i + cf_i(x)]_+ \nabla^2 f_i(x) + c\nabla f_i(x) \nabla f_i(x)^T$$

we may show that the generalized Hessian of $L_c(\cdot, y)$ is linearly generated by the matrices of the form

$$\nabla^2 f_0(x) + \sum_{i=1}^{n} \alpha_i H_c^i(x, y)$$

where the scalars α_i must be taken equal to zero, one, or in $[0, 1]$ following the case in which $[y_i + cf_i(x)]$ is negative, positive, or zero, respectively.

The Newton iteration of algorithm (A) then takes the form

$$x_{k+1} = x_k - \left[\nabla^2 f_0(x_k) + \sum_{i=1}^{n} \alpha_i H_c^i(x_k, y)\right]^{-1} \nabla_x L_c(x_k, y) \tag{27}$$

where it subsists an indetermination in the choice of the coefficients α_i when $[y_i + cf_i(x)] = 0$.

Another way to motivate iteration (27) is by linearization of the stationarity condition $\nabla_x L_c(x, y) = 0$ that can be equivalently written as

$$\nabla f_0(x) + \sum_{i=1}^{n} \lambda_i \nabla f_i(x) = 0, \qquad \lambda_i = [y_i + cf_i(x)]_+$$

and leads to the approximation

$$\nabla_x L_c(x_k, y) + \left[\nabla^2 f_0(x_k) + \sum_{i=1}^{n} \lambda_i^k \nabla^2 f_i(x_k)\right](x - x_k) + \sum_{i=1}^{n} (\lambda_i - \lambda_i^k) \nabla f_i(x_k) = 0$$

for determining x_{k+1} and $\lambda_i^{k+1} = [y_i + cf_i(x_{k+1})]_+$.

To estimate the difference $(\lambda_i^{k+1} - \lambda_i^k)$ we observe that when $[y_i + cf_i(x_k)]$ is positive or negative, the function $[y_i + cf_i(\cdot)]_+$ is differentiable at x_k and we obtain, respectively, $c\nabla f_i(x_k)(x_{k+1} - x_k)$ and zero as estimates for $(\lambda_i^{k+1} - \lambda_i^k)$. Similarly, in the case $[y_i + cf_i(x_k)] = 0$ we obtain the estimate $(\lambda_i^{k+1} - \lambda_i^k) \sim c[\nabla f_i(x_k)(x_{k+1} - x_k)]_+$.

These approximations lead to an iteration of the type of (27) with the α_i's chosen equal to one whenever $[y_i + cf_i(x_k)] > 0$ or when $[y_i + cf_i(x_k)] = 0$ but $\nabla f_i(x_k)(x_{k+1} - x_k) \geqq 0$, and equal to zero in the remaining cases. This is not a practical criteria, however,

since the point x_{k+1} is determined *after* the α_i's have been fixed. Nevertheless, this relation shows that the most interesting values for the α_i's are the extreme ones. A possible heuristic that might be suggested is to choose the α_i's more or less arbitrarily (take one of a previous iteration for instance) and eventually modify a posteriori the values of the α_i's that violate the previously established criteria.

REFERENCES

[1] R. ARAYA AND R. GORMAZ, *Problemas localmente Lipschitzianos en optimización*, thesis, Universidad de Chile, Santiago, Chile, 1979.

[2] J. P. AUBIN, *L'analyse Non Linéaire et ses Motivations Economiques*, Masson, Paris, 1984.

[3] ———, *Contingent derivatives of set-valued maps and existence of solutions to nonlinear inclusions and differential equations*, in Mathematical Analysis and Applications, L. Nachbin, ed., Academic Press, New York, 1981, pp. 159-229.

[4] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.

[5] A. AUSLENDER, *Optimisation: Méthodes Numériques*, Masson, Paris, 1976.

[6] ———, *On the differential properties of the support function of the ε-subdifferential of a convex function*, Math. Programming, 24 (1982), pp. 257-268.

[7] ———, *Stability in mathematical programming with nondifferentiable data*, SIAM J. Control Optim., 22 (1984), pp. 239-254.

[8] A. BEN-TAL AND J. ZOWE, *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Math. Programming Stud., 19 (1982), pp. 39-76.

[9] R. W. CHANEY, *Second order directional derivatives for nonsmooth functions*, preprint, Western Washington University, Bellingham, WA, 1985.

[10] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[11] R. COMINETTI, *Contribuciones al análisis no-diferenciable: derivadas generalizadas de primer y segundo orden*, thesis, Universidad de Chile, Santiago, Chile, 1986.

[12] ———, *Equivalence between the class of $\mathcal{C}^{1,1}$ and twice locally Lipschitzian function*, in Thesis, Université Blaise Pascal, 1989.

[13] R. COMINETTI AND R. CORREA, *Sur une dérivée du second ordre en analyse non différentiable*, C. R. Acad. Sci. Paris Sér. I, t.303 (1986), pp. 861-864.

[14] V. F. DEMYANOV, C. LEMARECHAL, AND J. ZOWE, *Approximation to a set valued mapping* I: *a proposal*, Appl. Math. Optim., 14 (1986), pp. 203-214.

[15] R. FLETCHER AND G. WATSON, *First and second order conditions for a class of nondifferentiable optimization problems*, Math. Programming, 18 (1980), pp. 291-307.

[16] J. B. HIRIART-URRUTY, *Mean value theorems for vector valued mappings in nonsmooth optimization*, Numer. Funct. Anal. Optim., 2 (1980), pp. 1-30.

[17] ———, *Approximating a second-order directional derivative for nonsmooth convex functions*, SIAM J. Control Optim., 20 (1982), pp. 783-807.

[18] ———, *Calculus rules on the approximate second-order directional derivative of a convex function*, SIAM J. Control Optim., 22 (1984), pp. 381-404.

[19] J. B. HIRIART-URRUTY, J.-J. STRODIOT, AND V. HIEN NGUYEN, *Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data*, Appl. Math. Optim., 11 (1984), pp. 43-56.

[20] J. B. HIRIART-URRUTY AND A. SEEGER, *Calculus rules on a new set valued second order derivative for convex functions*, preprint, 1985.

[21] J. B. HIRIART-URRUTY, *A new set valued second order derivative for convex functions*, in Mathematical Studies 129, J. B. Hiriart-Urruty, ed., North-Holland, Amsterdam, 1986, pp. 157-182.

[22] A. IOFFE, *Nonsmooth analysis: differential calculus of nondifferentiable mappings*, Trans. Amer. Math. Soc., 266 (1981), pp. 1-55.

[23] A. IOFFE AND T. MILOSZ, *Second order Lipschitzian functions and their application to optimization*, in preparation (1990).

[24] G. LEBOURG, *Generic differentiability of Lipschitzian functions*, Trans. Amer. Math. Soc., 256 (1979), pp. 125-144.

[25] C. LEMARECHAL AND E. NURMINSKI, *Sur la différentiabilité de la fonction d'appui du sous-différentiel approché*, C. R. Acad. Sci. Paris Sér. A, 290 (1980), pp. 855-858.

[26] T. MILOSZ, *Necessary and sufficient second order conditions for nonsmooth extremal problems*, thesis, Inst. Math. Polish Acad. Science, 1988. (In Russian.)

[27] D. PALLASHCKE AND P. RECHT, *On extensions of second order derivatives*, preprint, 1986.
[28] J. P. PENOT, *Differentiability of relations and differential stability of perturbed optimization problems*, SIAM J. Control Optim., 22 (1984), pp. 529–551.
[29] ———, *Generalized higher order derivatives and higher order optimality conditions*, preprint, Université de Pau, 1985.
[30] R. T. ROCKAFELLAR, *Maximal monotone relations and the second derivatives of nonsmooth functions*, Ann. Inst. Henri Poincaré, 2 (1985), pp. 167–184.
[31] A. SEEGER, *Analyse du second ordre de problèmes non différentiables*, thesis, Université Paul Sabatier, 1986.
[32] ———, *Second-order directional derivatives in parametric optimization problems*, Math. Oper. Res., 13 (1988), pp. 124–139.
[33] A. SHAPIRO, *Second-order derivatives of extremal valued functions and optimality conditions for semi-infinite programs*, Math. Oper. Res., 10 (1985), pp. 207–219.
[34] J. WARGA, *Second order necessary conditions in optimization*, SIAM J. Control Optim., 22 (1984), pp. 524–528.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 810-822, July 1990

003

GENERALIZED LINEAR-QUADRATIC PROBLEMS OF DETERMINISTIC AND STOCHASTIC OPTIMAL CONTROL IN DISCRETE TIME*

R. T. ROCKAFELLAR† AND R. J-B WETS‡

Abstract. Two fundamental classes of problems in large-scale linear and quadratic programming are described. Multistage problems covering a wide variety of models in dynamic programming and stochastic programming are represented in a new way. Strong properties of duality are revealed which support the development of iterative approximate techniques of solution in terms of saddlepoints. Optimality conditions are derived in a form that emphasizes the possibilities of decomposition.

Key words. discrete-time optimal control, dynamic programming, stochastic programming, large-scale linear-quadratic programming, intertemporal optimization, finite generation method

AMS(MOS) subject classifications. primary 19D40, 90C15; secondary 90C20

1. Introduction.

The importance of linear and quadratic programming problems is well appreciated in finite-dimensional optimization. Such problems serve as mathematical models in their own right and as subproblems solved within the context of general numerical methods of nonlinear programming. In optimal control, only a relatively small class of linear-quadratic problems has traditionally received much attention, however. A much more general class has recently been explored by Rockafellar [1] with the aim of opening up a wide domain for application of techniques of large-scale linear and quadratic programming, in particular the finite generation method of Rockafellar and Wets [2]–[4] that has been implemented in stochastic programming [5]. Central to this purpose is the development of flexible problem formulations for which there is a strong duality theory that represents optimal trajectories and controls in terms of saddlepoints of a "decomposable" Lagrangian.

In the present paper a discrete-time version of the deterministic models in [1] is investigated and corresponding results on optimality and duality are obtained. The formulations and results are then generalized to the stochastic case. The focus on discrete time is motivated by the computational possibilities already mentioned, so we do not hesitate to suppose also that the probability space for our stochastic version is discrete.

Our emphasis is on setting up a general framework for large-scale finite-dimensional linear-quadratic programming problems that reflect the special structure of optimal control. Besides being useful for numerical experimentation, such a framework may stimulate new applications, for instance, in areas like operations research and resource systems management, where inequality constraints occur that jointly involve states and controls. Although the task of clarifying the relationship between finite- and infinite-dimensional formulations is an important one, it is not the object of our efforts here.

In fact our discrete-time problems are more general than typical continuous-time problems in one respect: the dimensionality of the state and control vectors can vary with time. This feature is important in multistage modeling, where the decision structure in one period need not be the same as in another. The flexibility it provides allows us

* Received by the editors September 14, 1987; accepted for publication (in revised form) February 3, 1988. This research was supported in part by grants from the National Science Foundation and the Air Force Office of Scientific Research.

† Department of Mathematics, University of Washington, Seattle, Washington 98195.

‡ Department of Mathematics, University of California, Davis, California 95616.

to show that a much wider class of problems is covered by our format than might at first be imagined.

2. Generalized linear-quadratic programming. The control problems that will be formulated are based on a concept of generalized linear-quadratic programming explained fully in Rockafellar [1]. A problem fits this concept if it can be expressed in the form

$$(\mathcal{P}) \qquad \text{minimize } f(u) = \sup_{v \in V} J(u, v) \text{ over all } u \in U,$$

where U and V are polyhedral convex sets in $\mathbb{R}^k$ and $\mathbb{R}^l$, and J is a quadratic convex-concave function on $U \times V$, namely

$$J(u, v) = p \cdot u + \tfrac{1}{2} u \cdot Pu + q \cdot v - \tfrac{1}{2} v \cdot Qv - v \cdot Du, \tag{2.1}$$

where P and Q are symmetric and positive *semi*definite (possibly zero—we do not exclude "linear" when we say "quadratic," as we try to underline by sometimes using the term "linear-quadratic"). The problem dual to $(\mathcal{P})$ is then

$$(\mathcal{Q}) \qquad \text{maximize } g(v) = \inf_{u \in U} J(u, v) \text{ over all } v \in V.$$

Here $f(u)$ could be ∞ and $g(v)$ could be $-\infty$. We regard u as a feasible solution to $(\mathcal{P})$ only if $u \in U$ and $f(u) < \infty$; likewise, we regard v as a feasible solution to $(\mathcal{Q})$ only if $v \in V$ and $g(v) > -\infty$.

The expression of problems $(\mathcal{P})$ and $(\mathcal{Q})$ is facilitated by the notation

$$\rho_{V,Q}(r) = \sup_{v \in V} \{ r \cdot v - \tfrac{1}{2} v \cdot Qv \} \quad \text{for } r \in \mathbb{R}^l, \tag{2.2}$$

$$\rho_{U,P}(s) = \sup_{u \in U} \{ s \cdot u - \tfrac{1}{2} u \cdot Pu \} \quad \text{for } s \in \mathbb{R}^k. \tag{2.3}$$

Thus $\rho_{V,Q}$ is a function on $\mathbb{R}^l$ determined by the specification of a polyhedral convex set $V \subset \mathbb{R}^l$ and a symmetric positive semidefinite matrix $Q \in \mathbb{R}^{l \times l}$. It is in general "piecewise linear-quadratic" in a sense made precise in [1], and it may take on the value ∞. There are many special cases deserving of mention, but for these too one should consult [1]. Let it suffice to observe that when $0 \in V$, we have $\rho_{V,Q}(r) \geqq 0$ for all r, $\rho_{V,Q}(0) = 0$. Then $\rho_{V,Q}(r)$ can be interpreted as an expression that "monitors deviations of r from 0." Similarly for $\rho_{U,P}$.

In this notation our general problems can be written as

$$(\mathcal{P}) \qquad \text{minimize } p \cdot u + \tfrac{1}{2} u \cdot Pu + \rho_{V,Q}(q - Du) \text{ over } u \in U,$$

$$(\mathcal{Q}) \qquad \text{maximize } q \cdot v - \tfrac{1}{2} v \cdot Qv - \rho_{U,P}(D^* v - p) \text{ over } v \in V$$

(where the "*" signals the transpose matrix). In $(\mathcal{P})$, therefore, we have the possibility of linear constraints represented by the condition $u \in U$, and also an objective term that "monitors deviations of Du from q." This may be a penalty term that is zero for some kinds of deviations but positive for others. For example, if $V = \mathbb{R}^l_+$, $Q = 0$, we have

$$\rho_{V,Q}(q - Du) = \begin{cases} 0 & \text{if } Du \geqq q, \\ \infty & \text{if } Du \not\geqq q, \end{cases} \tag{2.4}$$

so that the ρ term in $(\mathcal{P})$ is a "sharp" representation of the constraint $Du \geqq q$. If at the same time one has $U = \mathbb{R}^k_+$, $P = 0$, then similarly

$$\rho_{U,P}(D^* v - p) = \begin{cases} 0 & \text{if } D^* v \leqq p, \\ \infty & \text{if } D^* v \not\leqq p. \end{cases} \tag{2.5}$$

In this case $(\mathscr{P})$ and $(\mathscr{Q})$ reduce to a canonical pair of linear programming problems in duality. See [1] for discussion of the rich possibilities that such ρ terms provide more generally in mathematical modeling.

The basic facts about the relationship between $(\mathscr{P})$ and $(\mathscr{Q})$ can be derived from the standard theory of linear and quadratic programming, specifically the duality theorem of Cottle [6] and the existence theorem of Frank and Wolfe [7].

THEOREM 2.1 (Rockafellar and Wets [3, Thm. 2]). *If either $(\mathscr{P})$ or $(\mathscr{Q})$ has finite optimal value, or if both problems have feasible solutions, then both optimal values are finite and equal, and both problems have optimal solutions. In this case a pair $(\bar{u}, \bar{v})$ is a saddlepoint of $J(u, v)$ relative to $u \in U$ and $v \in V$ if and only if $\bar{u}$ is an optimal solution to $(\mathscr{P})$ and $\bar{v}$ is an optimal solution to $(\mathscr{Q})$.*

3. Deterministic control model. We now want to formulate problems in this vein that belong to optimal control. The dynamical system we consider takes the form

$$\begin{aligned} &x_\tau = A_\tau x_{\tau-1} + B_\tau u_\tau + b_\tau \quad \text{for } \tau = 1, \cdots, T, \\ &x_0 = B_0 u_0 + b_0, \quad \text{where } u_\tau \in U_\tau \text{ for } \tau = 0, 1, \cdots, T. \end{aligned} \tag{3.1}$$

The vectors $u_\tau \in \mathbb{R}^{k_\tau}$ are controls, and the vectors $x_\tau \in \mathbb{R}^{n_\tau}$ are states (observe that dimensions can vary with τ). We write $u = (u_0, u_1, \cdots, u_T)$ and $x = (x_0, x_1, \cdots, x_T)$. Thus x is uniquely determined by u, and the transformation $u \mapsto x$ is affine. Note that u_0 serves as a supplementary parameter vector more than as a control vector in the usual dynamical sense.

The sets $U_\tau \subset \mathbb{R}^{k_\tau}$ are assumed to be polyhedral convex (nonempty). The matrices A_τ, B_τ and vectors b_τ are of appropriate dimension:

$$A_\tau \in \mathbb{R}^{n_\tau \times n_{\tau-1}}, \quad B_\tau \in \mathbb{R}^{n_\tau \times k_\tau}, \quad b_\tau \in \mathbb{R}^{n_\tau}.$$

(By taking $k_0 = 0$, one could eliminate u_0 from (3.1) and have $x_0 = b_0$.)

Our deterministic control problem is:

minimize subject to (3.1) the expression

$$\begin{aligned} (\mathscr{P}_{\text{det}}) \qquad f(u) = &\sum_{\tau=0}^{T} \left[p_\tau \cdot u_\tau + \frac{1}{2} u_\tau \cdot P_\tau u_\tau - c_{\tau+1} \cdot x_\tau \right] \\ &+ \sum_{\tau=1}^{T} \rho_{V_\tau, Q_\tau}(q_\tau - C_\tau x_{\tau-1} - D_\tau u_\tau) + \rho_{V_{T+1}, Q_{T+1}}(q_{T+1} - C_{T+1} x_T). \end{aligned}$$

Here V_τ is a polyhedral convex set (nonempty) in $\mathbb{R}^{l_\tau}$, and the matrices P_τ and Q_τ are symmetric and positive semidefinite. We have

$$\begin{gathered} P_\tau \in \mathbb{R}^{k_\tau \times k_\tau}, \quad Q_\tau \in \mathbb{R}^{l_\tau \times l_\tau}, \quad p_\tau \in \mathbb{R}^{k_\tau}, \quad q_\tau \in \mathbb{R}^{l_\tau}, \\ c_\tau \in \mathbb{R}^{n_{\tau-1}}, \quad C_\tau \in \mathbb{R}^{l_\tau \times n_{\tau-1}}, \quad D_\tau \in \mathbb{R}^{l_\tau \times k_\tau}. \end{gathered}$$

In this notation the elements A_τ and D_τ are defined only for $\tau = 1, \cdots, T$, but B_τ, b_τ, P_τ, p_τ, are defined for $\tau = 0, 1, \cdots, T$ and C_τ, c_τ, Q_τ, q_τ for $\tau = 1, \cdots, T, T+1$.

For the problem that will turn out to be dual to $(\mathscr{P}_{\text{det}})$, the dynamical system goes backward in time:

$$\begin{aligned} &y_\tau = A_\tau^* y_{\tau+1} + C_\tau^* v_\tau + c_\tau \quad \text{for } \tau = 1, \cdots, T, \\ &y_{T+1} = C_{T+1}^* v_{T+1} + c_{T+1}, \quad \text{where } v_\tau \in V_\tau \text{ for } \tau = 1, \cdots, T, T+1. \end{aligned} \tag{3.2}$$

The vectors $v_\tau \in \mathbb{R}^{l_\tau}$ are the dual controls, and the vectors $y_\tau \in \mathbb{R}^{n_{\tau-1}}$ are the dual states.

We write

$$v=(v_1,\cdots,v_T,v_{T+1}) \quad \text{and} \quad y=(y_1,\cdots,y_T,y_{T+1}).$$

The dual problem then is

$(\mathcal{Q}_{\text{det}})$ maximize subject to (3.2) the expression

$$g(v)=\sum_{\tau=1}^{T+1}\left[q_\tau\cdot v_\tau-\frac{1}{2}v_\tau\cdot Q_\tau v_\tau-b_{\tau-1}\cdot y_\tau\right]$$
$$-\sum_{\tau=1}^{T}\rho_{U_\tau,P_\tau}(B_\tau^*y_{\tau+1}-D_\tau^*v_\tau-p_\tau)-\rho_{U_0,P_0}(B_0^*y_1-p_0).$$

In this formula y is the trajectory uniquely determined from v by (3.2).

PROPOSITION 3.1. *Suppose x corresponds to u by* (3.1), *and y to v by* (3.2). *Then*

$$\sum_{\tau=0}^{T}y_{\tau+1}\cdot[B_\tau u_\tau+b_\tau]=\sum_{\tau=1}^{T+1}x_{\tau-1}\cdot[C_\tau^*v_\tau+c_\tau]. \tag{3.3}$$

Proof. In view of the relations (3.1) the left side of (3.3) can be written as

$$y_1\cdot x_0+\sum_{\tau=1}^{T}y_{\tau+1}[x_\tau-A_\tau x_{\tau-1}]=y_1\cdot x_0+y_2\cdot x_1+\cdots+y_{T+1}\cdot x_T-\sum_{\tau=1}^{T}x_{\tau-1}\cdot A_\tau^*y_{\tau+1}.$$

Likewise from (3.2) the right side becomes

$$x_T\cdot y_{T+1}+\sum_{\tau=1}^{T}x_{\tau-1}\cdot[y_\tau-A_\tau^*y_{\tau+1}]=y_1\cdot x_0+y_2\cdot x_1+\cdots+y_{T+1}\cdot x_T-\sum_{\tau=1}^{T}x_{\tau-1}\cdot A_\tau^*y_{\tau+1}.$$

Thus the two sides are equal, as claimed. □

PROPOSITION 3.2. *Let $U=U_0\times\cdots\times U_T$ and $V=V_1\times\cdots\times V_{T+1}$, and for $u\in U$ and $v\in V$ define*

$$J(u,v)=\sum_{\tau=0}^{T}\left(p_\tau\cdot u_\tau+\frac{1}{2}u_\tau\cdot P_\tau u_\tau\right)+\sum_{\tau=1}^{T+1}\left(q_\tau\cdot v_\tau-\frac{1}{2}v_\tau\cdot Q_\tau v_\tau\right)$$
$$-\sum_{\tau=1}^{T}v_\tau\cdot D_\tau u_\tau-[u,v], \tag{3.4}$$

where $[u,v]$ denotes the common value of the expression in (3.3).

Then U and V are polyhedral convex sets, and J is a quadratic convex-concave function.

Proof. This is immediate from our assumptions and the fact that the expression $[u,v]$ is affine in u and v separately. □

THEOREM 3.3. *The deterministic optimal control problems $(\mathcal{P}_{\text{det}})$ and $(\mathcal{Q}_{\text{det}})$ are the primal and dual problems of generalized linear-quadratic programming associated with the U, V, and J in Proposition* 3.2. *In particular, the assertions of Theorem* 2.1 *are valid for $(\mathcal{P}_{\text{det}})$ and $(\mathcal{Q}_{\text{det}})$.*

Proof. We need only show that the expressions $f(u)$ and $g(v)$ in $(\mathcal{P}_{\text{det}})$ and $(\mathcal{Q}_{\text{det}})$ arise according to the pattern in the general problems $(\mathcal{P})$ and $(\mathcal{Q})$ of § 1. First, using for $[u,v]$ in (3.4) the right-hand expression in (3.3), we write

$$J(u,v)=\sum_{\tau=0}^{T}\left(p_\tau\cdot u_\tau+\frac{1}{2}u_\tau\cdot P_\tau u_\tau\right)-\sum_{\tau=1}^{T+1}c_\tau\cdot x_{\tau-1}$$
$$+\sum_{\tau=1}^{T}\left([q_\tau-C_\tau x_{\tau-1}-D_\tau u_\tau]\cdot v_\tau-\frac{1}{2}v_\tau\cdot Q_\tau v_\tau\right) \tag{3.5}$$
$$+\left([q_{T+1}-C_{T+1}x_T]\cdot v_{T+1}-\frac{1}{2}v_{T+1}\cdot Q_{T+1}v_{T+1}\right).$$

The maximization of this over all $v \in V$ reduces to a separate maximization with respect to each of the components v_τ of v. Since by definition

$$\sup_{v_\tau \in V_\tau} \{[q_\tau - C_\tau x_{\tau-1} - D_\tau u_\tau] \cdot v_\tau - \tfrac{1}{2} v_\tau \cdot Q_\tau v_\tau\} = \rho_{V_\tau, Q_\tau}(q_\tau - C_\tau x_{\tau-1} - D_\tau u_\tau)$$

and

$$\sup_{v_{T+1} \in V_{T+1}} \{[q_{T+1} - C_{T+1} x_T] \cdot v_{T+1} - \tfrac{1}{2} v_{T+1} \cdot Q_{T+1} v_{T+1}\} = \rho_{V_{T+1}, Q_{T+1}}(q_{T+1} - C_{T+1} x_T),$$

we conclude that $\sup_{v \in V} J(u, v)$ is the $f(u)$ in $(\mathcal{P}_{\text{det}})$.

Next, using for $[u, v]$ the left-hand expression in (3.3), we write

$$\begin{aligned} J(u, v) = &\sum_{\tau=1}^{T+1} \left(q_\tau \cdot v_\tau - \frac{1}{2} v_t \cdot Q_\tau v_\tau \right) - \sum_{\tau=0}^{T} b_\tau \cdot y_{\tau+1} \\ &- \sum_{\tau=1}^{T} \left([B_\tau^* y_{\tau+1} + D_\tau^* v_\tau - p_\tau] \cdot u_\tau - \frac{1}{2} u_\tau \cdot P_\tau u_\tau \right) \\ &- \left([B_0^* y_1 - p_0] \cdot u_0 - \frac{1}{2} u_0 \cdot P_0 u_0 \right). \end{aligned} \tag{3.6}$$

The minimization of this over all $u \in U$ reduces similarly to a separate minimization with respect to each of the components u_τ. We know that

$$\sup_{u_\tau \in U_\tau} \{[B_\tau^* y_{\tau+1} + D_\tau^* v_\tau - p_\tau] \cdot u_\tau - \tfrac{1}{2} u_\tau \cdot P_\tau u_\tau\} = \rho_{u_\tau, P_\tau}(B_\tau^* y_{\tau+1} + D_\tau^* v_\tau - p_\tau)$$

and

$$\sup_{u_0 \in U_0} \{[B_0^* y_1 - p_0] \cdot u_0 - \tfrac{1}{2} u_0 \cdot P_0 u_0\} = \rho_{U_0, P_0}(B_0^* y_1 - p_0).$$

We conclude that $\inf_{u \in U} J(u, v)$ is the $g(v)$ in $(\mathcal{Q}_{\text{det}})$. □

The proof of Theorem 3.3 reveals an important simplifying feature of our minimax representation of $(\mathcal{P}_{\text{det}})$ and $(\mathcal{Q}_{\text{det}})$. We state it as follows.

THEOREM 3.4. *For the U, V, and J in Theorem* 3.3 *one has the following decomposability properties for separate minimization in u or maximization in v. Here* $\bar{u}$ *and* $\bar{v}$ *are elements of U and V, and* $\bar{x}$ *and* $\bar{y}$ *the corresponding trajectories.*

(a) $\tilde{u} \in \operatorname{argmin}_{u \in U} J(u, \bar{v})$ *if and only if*

$$\tilde{u}_\tau \in \partial \rho_{U_\tau, P_\tau}(B_\tau^* \bar{y}_{\tau+1} + D_\tau^* \bar{v}_\tau - p_\tau) = \operatorname*{argmax}_{u_\tau \in U_\tau} \{[B_\tau^* \bar{y}_{\tau+1} + D_\tau^* \bar{v}_\tau - p_\tau] \cdot u_\tau - \tfrac{1}{2} u_\tau \cdot P_\tau u_\tau\}$$

for $\tau = 1, \cdots, T$, *and*

$$\tilde{u}_0 \in \partial \rho_{U_0, P_0}(B_0^* \bar{y}_1 - p_0) = \operatorname*{argmax}_{u_0 \in U_0} \{[B_0^* \bar{y}_1 - p_0] - \tfrac{1}{2} u_0 \cdot P_0 u_0\}.$$

(b) $\tilde{v} \in \operatorname{argmax}_{v \in V} J(\bar{u}, v)$ *if and only if*

$$\tilde{v}_\tau \in \partial \rho_{V_\tau, Q_\tau}(q_\tau - C_\tau \bar{x}_{\tau-1} - D_\tau \bar{u}_\tau) = \operatorname*{argmax}_{v_\tau \in V_\tau} \{[q_\tau - C_\tau \bar{x}_{\tau-1} - D_\tau \bar{u}_\tau] \cdot v_\tau - \tfrac{1}{2} v_\tau \cdot Q_\tau v_\tau\}$$

for $\tau = 1, \cdots, T$, *and*

$$\tilde{v}_{T+1} \in \partial \rho_{V_{T+1}, Q_{T+1}}(q_{T+1} - C_{T+1} \bar{x}_T) = \operatorname*{argmax}_{v_{T+1} \in V_{T+1}} \{[q_{T+1} - C_{T+1} \bar{x}_t] \cdot v_{T+1} - \tfrac{1}{2} v_{T+1} \cdot Q_{T+1} v_{T+1}\}.$$

Proof. The formulas in terms of "argmax" are justified by the calculations in the proof of Theorem 3.3. The question that remains is whether the "argmax" sets are

truly the same as the indicated subgradient sets. This is answered by the observation that in the notation (2.2) we have $\rho_{V,Q} = \theta^*_{V,Q}$ (convex conjugate), where

$$\theta_{V,Q}(v) = \begin{cases} \frac{1}{2} v \cdot Qv & \text{if } v \in V, \\ \infty & \text{if } v \notin V. \end{cases} \tag{3.7}$$

Inasmuch as $\theta_{V,Q}$ is a closed proper convex function, we also have $\theta_{V,Q} = \rho^*_{V,Q}$ and

$$\partial \rho_{V,Q}(r) = \operatorname*{argmax}_{v \in \mathbb{R}^l} \{ r \cdot v - \theta_{V,Q}(v) \} \tag{3.8}$$

by the basic rules of convex analysis [8, Thm. 12.2]. When this is applied to the pairs V_τ, Q_τ, and U_τ, P_τ, in place of V, Q, we reach our desired conclusion. □

The significance of the formulas in Theorem 3.4 lies in their potential use in iterative methods for solving $(\mathcal{P}_{\text{det}})$ and $(\mathcal{Q}_{\text{det}})$ when the dimensions

$$k = \sum_{\tau=0}^{T} k_\tau \quad \text{and} \quad l = \sum_{\tau=1}^{T+1} l_\tau \tag{3.9}$$

of the vectors $u = (u_0, u_1, \cdots, u_T)$ and $v = (v_1, \cdots, v_T, v_{T+1})$ are large. The dimensions may be expected to be large if T is large, as of course would happen in particular in taking $(\mathcal{P}_{\text{det}})$ and $(\mathcal{Q}_{\text{det}})$ to be discrete-time approximations to continuous-time control problems such as the ones studied in [1]. In the presence of high dimensions, it may be impossible or inexpedient to solve $(\mathcal{P}_{\text{det}})$ and $(\mathcal{Q}_{\text{det}})$ directly by reducing them to ordinary quadratic programming problems in duality and applying a typical finitely-terminating quadratic programming code (as would be possible in principle in a manner explained in Rockafellar and Wets [3, § 2]).

An alternative approach in that case is the exploration of methods that determine approximate solutions to $(\mathcal{P}_{\text{det}})$ and $(\mathcal{Q}_{\text{det}})$ by calculating a sequence of approximate saddlepoints $(\bar{u}^\nu, \bar{v}^\nu)$ of J on $U \times V$ for $\nu = 1, 2, \cdots$, as suggested by the characterization of optimality in Theorem 3.4. In any such method the ability to calculate

$$f(\bar{u}^\nu) = \max_{v \in V} J(\bar{u}^\nu, v) \quad \text{and} \quad \tilde{v}^\nu \in \operatorname*{argmax}_{v \in V} J(\bar{u}^\nu, v) \tag{3.10}$$

as well as

$$g(\bar{v}^\nu) = \min_{u \in U} J(u, \bar{v}^\nu) \quad \text{and} \quad \tilde{u}^\nu \in \operatorname*{argmin}_{u \in U} J(u, \bar{v}^\nu) \tag{3.11}$$

is crucial in producing primal and dual bounds that tell how far $\bar{u}^\nu$ and $\bar{v}^\nu$ are from optimality and as input to possible schemes for updating $(\bar{u}^\nu, \bar{v}^\nu)$ to $(\bar{u}^{\nu+1}, \bar{v}^{\nu+1})$. Theorem 3.4 says that the calculations in (3.10) and (3.11) can feasibly be carried out in terms of solving a collection of low-dimensional quadratic programming subproblems indexed by τ. Moreover these subproblems can even be solved in "closed form," i.e., without applying a quadratic programming code, if the functions ρ_{V_τ, Q_τ} and ρ_{U_τ, P_τ} have sufficiently simple expressions that allow the use of subgradient formulas directly.

The subgradient formulas are readily usable, for example, in the completely decomposable case where U_τ and V_τ are boxes (products of closed intervals, e.g., orthants) and P_τ and Q_τ are diagonal. Indeed, if P_τ and Q_τ are nonsingular, the subgradients reduce to gradients given by very elementary expressions.

THEOREM 3.5. *Consider a control pair $\bar{u}$, $\bar{v}$, and the corresponding trajectories $\bar{x}$ and $\bar{y}$ determined by* (3.1) *and* (3.2). *Define*

$$\begin{aligned} &\bar{p}_\tau = p_\tau - B^*_\tau \bar{y}_{\tau+1} \quad \textit{for } \tau = 0, 1, \cdots, T, \\ &\textit{and} \\ &\bar{q}_\tau = q_\tau - C_\tau \bar{x}_{T-1} \quad \textit{for } \tau = 1, \cdots, T, T+1. \end{aligned} \tag{3.12}$$

Let $(\bar{\mathcal{P}}_\tau)$ *and* $(\bar{\mathcal{Q}}_\tau)$ *for* $\tau = 1, \cdots, T$ *denote the primal and dual problems of generalized linear-quadratic programming associated with*

$$J_\tau(u_\tau, v_\tau) = \bar{p}_\tau u_\tau + \tfrac{1}{2} u_\tau \cdot P_\tau u_\tau + \bar{q}_\tau \cdot v_\tau - \tfrac{1}{2} v_\tau \cdot Q_\tau v_\tau - v_\tau \cdot D_\tau u_\tau \tag{3.13}$$

on $U_\tau \times V_\tau$, *namely,*

$(\bar{\mathcal{P}}_\tau)$ *minimize* $\bar{p}_\tau \cdot u_\tau + \tfrac{1}{2} u_\tau \cdot P_\tau u_\tau + \rho_{V_\tau, Q_\tau}(\bar{q}_\tau - D_\tau u_\tau)$ *over* $u_\tau \in U_\tau$,

$(\bar{\mathcal{Q}}_\tau)$ *maximize* $\bar{q}_\tau \cdot v_\tau - \tfrac{1}{2} v_\tau \cdot Q_\tau v_\tau - \rho_{U_\tau, P_\tau}(D_\tau^* v_\tau - \bar{p}_\tau)$ *over* $v_\tau \in V_\tau$,

and consider also the problems

$(\bar{\mathcal{P}}_0)$ *minimize* $\bar{p}_0 \cdot u_0 + \tfrac{1}{2} u_0 \cdot P_0 u_0$ *over* $u_0 \in U_0$,

$(\bar{\mathcal{Q}}_{T+1})$ *maximize* $\bar{q}_{T+1} \cdot v_{T+1} - \tfrac{1}{2} v_{T+1} \cdot Q_{T+1} v_{T+1}$ *over* $v_{T+1} \in V_{T+1}$.

Then a necessary and sufficient condition for $\bar{u}$ *and* $\bar{v}$ *to be optimal solutions to the control problems* $(\mathcal{P}_{\text{det}})$ *and* $(\mathcal{Q}_{\text{det}})$, *respectively, is that* $\bar{u}_\tau$ *should be an optimal solution to the subproblem* $(\bar{\mathcal{P}}_\tau)$ *for* $\tau = 0, 1, \cdots, T$, *and* $\bar{v}_\tau$ *should be an optimal solution to the subproblem* $(\bar{\mathcal{Q}}_\tau)$ *for* $\tau = 1, \cdots, T, T+1$.

Proof. We know from Theorem 3.3 that a necessary and sufficient condition for the optimality of $\bar{u}$ and $\bar{v}$ in $(\mathcal{P}_{\text{det}})$ and $(\mathcal{Q}_{\text{det}})$ is the saddlepoint relation

$$\bar{u} \in \operatorname*{argmin}_{u \in U} J(u, \bar{v}) \quad \text{and} \quad \bar{v} \in \operatorname*{argmax}_{v \in V} J(\bar{u}, v).$$

Furthermore, this reduces to having the argmax conditions in Theorem 3.4 hold for $\tilde{u} = \bar{u}$ and $\tilde{v} = \bar{v}$. These conditions in turn are equivalent to

$$\bar{u}_\tau \in \operatorname*{argmin}_{u_\tau \in U_\tau} J_\tau(u_\tau, \bar{v}_\tau) \quad \text{for } \tau = 1, \cdots, T,$$

$$\bar{u}_0 \in \operatorname*{argmin}_{u_0 \in U_0} \{\bar{p}_0 \cdot u_0 + \tfrac{1}{2} u_0 \cdot P_0 u_0\},$$

and

$$\bar{v}_\tau \in \operatorname*{argmax}_{v_\tau \in V_\tau} J_\tau(\bar{u}_\tau, v_\tau) \quad \text{for } \tau = 1, \cdots, T,$$

$$\bar{v}_{T+1} \in \operatorname*{argmax}_{v_{T+1} \in V_{T+1}} \{\bar{q}_{T+1} \cdot v_{T+1} - \tfrac{1}{2} v_{T+1} \cdot Q_{T+1} v_{T+1}\}.$$

The latter mean that $\bar{u}_0$ is optimal for $(\mathcal{P}_0)$, $\bar{v}_{T+1}$ is optimal for $(\mathcal{Q}_{T+1})$, and $(\bar{u}_\tau, \bar{v}_\tau)$ is a saddlepoint of $J_\tau(u_\tau, v_\tau)$ relative to $u_\tau \in U_\tau$ and $v_\tau \in V_\tau$ for $\tau = 1, \cdots, T$. This saddlepoint condition is equivalent by Theorem 2.1 to $\bar{u}_\tau$ and $\bar{v}_\tau$ being optimal solutions to the primal and dual subproblems $(\bar{\mathcal{P}}_\tau)$ and $(\bar{\mathcal{Q}}_\tau)$. □

Optimality conditions of the kind in Theorem 3.5 were developed for continuous-time problems in Rockafellar [1]. They resemble conditions first detected in a special setting known as "continuous linear programming" by Grinold [9].

Besides being of interest in the study of what optimality might mean in a particular application modeled directly in terms of $(\mathcal{P}_{\text{det}})$ and (Q_{det}), the conditions in Theorem 3.5, like those in Theorem 3.4, have import for computations. Having arrived at a control pair $(\bar{u}^\nu, \bar{v}^\nu)$ and associated trajectories $(\bar{x}^\nu, \bar{y}^\nu)$ in some iteration ν of a numerical method, one can construct a new pair $(u^\nu, v^\nu) \in U \times V$ by taking u_τ^ν to be an optimal solution to $(\bar{\mathcal{P}}_\tau^\nu)$ for $\tau = 0, 1, \cdots, T$ and v_τ^ν an optimal solution to $(\bar{\mathcal{Q}}_\tau^\nu)$ for $\tau = 1, \cdots, T, T+1$, where $(\bar{\mathcal{P}}_\tau^\nu)$ and $(\bar{\mathcal{Q}}_\tau^\nu)$ are the subproblems corresponding to $\bar{u}^\nu$ and $\bar{v}^\nu$ in the sense of Theorem 3.5. Then u^ν and v^ν generate new trajectories x^ν and y^ν that may be compared with $\bar{x}^\nu$ and $\bar{y}^\nu$, and so forth. This procedure, like the one

described after Theorem 3.4, provides another tool that might, after suitable elaboration, be used constructively in the generation of a sequence of approximate saddle points.

4. Stochastic control model. The probability space we work with in this paper is simply a finite set Ω, for reasons given in § 1. The probability associated with an element $\omega \in \Omega$ is $\pi_\omega \geqq 0$; we have $\sum_{\omega \in \Omega} \pi_\omega = 1$. The vectors, matrices, and sets introduced in the formulation of our deterministic problems persist notationally in the stochastic problems, but all are now treated as (potentially) random variables. Thus, for example, p_τ now denotes a mapping $\omega \mapsto p_{\omega\tau} \in \mathbb{R}^{k_\tau}$ rather than necessarily just a single vector. Likewise P_τ is a matrix-valued mapping $\omega \mapsto P_{\omega\tau}$, and U_τ is a set-valued mapping $\omega \mapsto U_{\omega\tau}$. In line with our earlier assumptions, we suppose that $P_{\omega\tau}$ and $Q_{\omega\tau}$ are *positive semidefinite* (symmetric), and $U_{\omega\tau}$ and $V_{\omega\tau}$ are *polyhedral convex* (nonempty). The expectation of a random variable such as p_τ is

$$E\{p_\tau\} = E_\omega\{p_{\omega\tau}\} := \sum_{\omega \in \Omega} \pi_\omega p_{\omega\tau}.$$

The information available to the decision-making process at time τ is modeled by the specification of a (finite) field $\mathcal{G}_\tau$ of subsets of Ω for $\tau = 0, 1, \cdots, T, T+1$. The fields $\mathcal{G}_\tau$ may differ from the complete information fields $\mathcal{F}_\tau$, and no particular relation between them is presupposed, although the case where the $\mathcal{G}_\tau$'s are increasing with $\mathcal{G}_\tau$ contained in $\mathcal{F}_\tau$ is, for instance, an important one. More will be said about this after the statement of our primal and dual problems. We assume that

$$U_\tau,\ V_\tau,\ p_\tau,\ P_\tau,\ q_\tau,\ Q_\tau, \text{ and } D_\tau \text{ are } \mathcal{G}_\tau\text{-measurable,} \tag{4.1}$$

but in general do *not* place this restriction on A_τ, B_τ, C_τ, b_τ, or c_τ. Trivially the latter are measurable with respect to the underlying field $\mathcal{F}$ of complete information, comprised here of *all* the subsets of Ω.

Because $\mathcal{G}_\tau$ is a finite collection of subsets of Ω, the notion of $\mathcal{G}_\tau$-measurability has an especially simple representation for our purposes. Let $\mathcal{A}_\tau$ denote the subcollection of $\mathcal{G}_\tau$ consisting of all *$\mathcal{G}_\tau$-atoms*, i.e., nonempty $\mathcal{G}_\tau$-measurable sets that do not properly include any other nonempty $\mathcal{G}_\tau$-measurable set. Such atoms are mutually disjoint. A set is $\mathcal{G}_\tau$-measurable if and only if it is a union of $\mathcal{G}_\tau$-atoms. Thus there is a one-to-one correspondence between $\mathcal{G}_\tau$-measurable sets in Ω and sets of $\mathcal{G}_\tau$-atoms, i.e., subsets of $\mathcal{A}_\tau$. A function is $\mathcal{G}_\tau$-measurable if and only if it is constant relative to every $\mathcal{G}_\tau$-atom. Each $\mathcal{G}_\tau$-measurable function can in this way be identified uniquely with a function on $\mathcal{A}_\tau$ rather than on Ω. We can indicate this notationally, when we wish to, by writing $p_{\alpha\tau}$ for $\alpha \in \mathcal{A}_\tau$ to denote the common value that $p_{\omega\tau}$ has for all $\omega \in \alpha$ when p is $\mathcal{G}_\tau$-measurable. (Obviously Ω itself in this setting might be identified with the set of atoms of some finite field of information chosen within a larger, possibly "continuous" probability space by some kind of approximation. We do not go into this matter here.)

Conditional expectation with respect to $\mathcal{G}_\tau$ is denoted by $E^{\mathcal{G}_\tau}$. This can be viewed in the present setting as the linear transformation that takes a random variable such as B_τ and redefines it to have a constant value on each $\mathcal{G}_\tau$-atom $\alpha \in \mathcal{A}_\tau$, that value being, of course, the "weighted average"

$$\left[\sum_{\omega \in \alpha} \pi_\omega B_{\omega\tau}\right] \Big/ \left[\sum_{\omega \in \alpha} \pi_\omega\right].$$

The stochastic dynamical systems for our primal and dual problems are taken again to have the forms (3.1) and (3.2), but with all elements now interpreted as

(potentially) random, and with the restriction that

$$u_\tau \text{ is } \mathcal{G}_\tau\text{-measurable}, \tag{4.2}$$

$$v_\tau \text{ is } \mathcal{G}_\tau\text{-measurable}. \tag{4.3}$$

The condition $u_\tau \in U_\tau$ in (3.1) is interpreted to mean that $u_{\omega\tau} \in U_{\omega\tau}$ for all $\omega \in \Omega$, and similarly for $v_\tau \in V_\tau$. Our primal problem of stochastic control is

$$(\mathcal{P}_{\text{sto}}) \quad \begin{aligned} &\text{minimize subject to (3.1) and (4.2) the function} \\ &f(u) = \sum_{\tau=0}^{T} E\left\{p_\tau \cdot u_\tau + \frac{1}{2} u_\tau \cdot P_\tau u_\tau\right\} - \sum_{\tau=1}^{T+1} E\{c_\tau \cdot x_{\tau-1}\} \\ &\quad + \sum_{\tau=1}^{T} E\{\rho_{V_\tau, Q_\tau}(q_\tau - E^{\mathcal{G}_\tau}\{C_\tau x_{\tau-1}\} - D_\tau u_\tau)\} \\ &\quad + E\{\rho_{V_{T+1}, Q_{T+1}}(q_{T+1} - E^{\mathcal{G}_{T+1}}\{C_{T+1} x_T\})\}. \end{aligned}$$

The corresponding dual problem is

$$(\mathcal{Q}_{\text{sto}}) \quad \begin{aligned} &\text{maximize subject to (3.2) and (4.3) the function} \\ &g(v) = \sum_{\tau=1}^{T+1} E\left\{q_\tau \cdot v_\tau - \frac{1}{2} v_\tau \cdot Q_\tau v_\tau\right\} - \sum_{\tau=1}^{T} E\{b_\tau \cdot y_{\tau+1}\} \\ &\quad - \sum_{\tau=1}^{T} E\{\rho_{U_\tau, P_\tau}(E^{\mathcal{G}_\tau}\{B_\tau^* y_{\tau+1}\} + D_\tau^* v_\tau - p_\tau)\} \\ &\quad - E\{\rho_{U_0, P_0}(E^{\mathcal{G}_0}\{B_0^* y_1 - p_0\})\}. \end{aligned}$$

Here ρ_{V_τ, Q_τ} and ρ_{U_τ, P_τ} are "random functions" that depend $\mathcal{G}_\tau$-measurably on $\omega \in \Omega$ by virtue of (4.1). The random variables

$$\xi_\tau := E^{\mathcal{G}_\tau}\{C_\tau x_{\tau-1}\} \quad \text{and} \quad \eta_\tau := E^{\mathcal{G}_\tau}\{B_\tau^* y_{\tau+1}\} \tag{4.4}$$

are $\mathcal{G}_\tau$-measurable too, of course, so the arguments to which ρ_{V_τ, Q_τ} and ρ_{U_τ, P_τ} are applied are always $\mathcal{G}_\tau$-measurable. The ρ terms at time τ thus monitor "constraint expressions" based solely on the information available to the decision maker at time τ. Note from the dynamics that $\xi_{\omega\tau}$ depends affinely on $u_{\omega 0}, \cdots, u_{\omega,\tau-1}$, whereas $\eta_{\omega\tau}$ depends affinely on $v_{\omega,\tau+1}, \cdots, v_{\omega,T+1}$.

In order to appreciate the generality of problem (P_{sto}) it is important, especially for readers accustomed to the traditional approach to stochastic control, to understand the nature of the information structure that is adopted. This structure, which is typical of the literature on stochastic programming, has sometimes been interpreted narrowly as excluding models where the information on which decisions can be based is generated by observations that may be influenced by previous control decisions, cf. the comments of Bertsekas and Shreve [13, pp. 10–11]. Such is not actually the case when measurability requirements are referred to a single underlying space, as we shall explain. Thus the specification of the information field $\mathcal{G}_\tau$ as independent of $u_0, u_1, \cdots, u_{\tau-1}$, should *not* be taken to mean, for instance, that in choosing u_τ we are unable to respond to complete or partial observations of the states $x_0, x_1, \cdots, x_{\tau-1}$, inasmuch as those states are generally random variables whose distributions depend on $u_0, u_1, \cdots, u_{\tau-1}$.

The crucial distinction is that of controls u_τ seen directly as functions on the space Ω, rather than controls represented in a feedback mode as functions of past observations and expressible only in a secondary way, through composition, as functions on Ω. The feedback mode of representation, while conceptually very appealing, can be a handicap

in our opinion when imposed right from the beginning in the problem formulation. We prefer to proceed at first without it and to recover feedback laws later from optimality conditions, if desired.

Let us imagine, to make this more explicit, that at each time $\tau = 0, 1, \cdots, N$ an observation $z_\tau \in \mathbb{R}^{m_\tau}$ is made before the control decision u_τ is chosen. Of course z_τ is a random variable whose distribution is given by a probability measure μ_τ on $\mathbb{R}^{m_\tau}$, which in general might depend on the controls $u_0, u_1, \cdots, u_{\tau-1}$. Let us suppose that the only information available for the selection of u_τ is the sequence $z_0, z_1, \cdots, z_\tau$. In stochastic control it is common to express this requirement by taking u_τ to be a function of $z_0, z_1, \cdots, z_\tau$, i.e., as a function of a random argument in $\mathbb{R}^{m_0} \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_\tau}$. What we propose instead is to handle $z_0, z_1, \cdots, z_\tau$ as functions defined on the underlying probability space Ω and take u_τ to be a function on Ω that is measurable with respect to the σ-field generated by $z_0, z_1, \cdots, z_\tau$; it is this field that should be identified with $\mathcal{G}_\tau$ in our model. (We have assumed in this paper that Ω is a finite, discrete set, but the idea under consideration applies more generally.) This condition is tantamount to the requirement that u_τ be *representable* by composition of $(z_0, z_1, \cdots, z_\tau)$ with some mapping into $\mathbb{R}^{k_\tau}$ from the probability space in $\mathbb{R}^{m_0} \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_\tau}$ induced by these random variables, but it leaves the particular representation open to later investigation.

The advantage to our approach in this setting is that the field $\mathcal{G}_\tau$ may well be independent of $u_0, u_1, \cdots, u_{\tau-1}$, even though the distribution of $(z_0, z_1, \cdots, z_\tau)$ might not. To this extent we are able to make use of properties of convexity and duality that otherwise could be overlooked.

Before we return to the characterization of optimal controls and trajectories, let us also note that because we allow the dimensionality of the state and control vectors to vary over time, our model also includes classical multistage recourse problems. Suppose that the equations (3.1) have the special form

$$x_\tau = \begin{bmatrix} I \\ 0 \end{bmatrix} x_{\tau-1} + \begin{bmatrix} 0 \\ I \end{bmatrix} u_\tau \quad \text{for } \tau = 1, \cdots, T,$$

$$x_0 = u_0,$$

where the identity matrices I and zero matrices 0 are of the appropriate dimensions. Then

$$x_0 = u_0, \quad x_1 = \begin{bmatrix} u_0 \\ u_1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix}, \quad \text{etc.}$$

Thus x_τ is the "memory" of all decisions up through time τ. Assuming that $\mathcal{G}_{\tau-1} \subset \mathcal{G}_\tau$, we get x_τ, like u_τ, to be $\mathcal{G}_\tau$-measurable. Then in $(\mathcal{P}_{\text{sto}})$ the term

$$q_\tau - E^{\mathcal{G}_\tau}\{C_\tau x_{\tau-1}\} - D_\tau u_\tau$$

represents a general affine expression in $u_0, u_1, \cdots, u_\tau$. When ρ_{V_τ, Q_τ} is of the type (2.4), we can rewrite $(\mathcal{P}_{\text{sto}})$ in terms of linear constraints and a quadratic objective involving only the control variables $u_0, u_1, \cdots, u_T$. This problem, with its block angular structure, is in the usual format for the multistage stochastic programs with recourse; see [11] or [12], for example.

Problem $(\mathcal{P}_{\text{sto}})$ revolves around the choice of the random variable $u = (u_0, u_1, \cdots, u_T)$, which can be regarded as a function from Ω to $\mathbb{R}^{k_0} \times \cdots \times \mathbb{R}^{k_T}$ and therefore as an element of the finite-dimensional vector space consisting of all such functions. The dimension of this space may be very large indeed just from the size of Ω and possibly T, even if $k_0, \cdots, k_T$ are themselves relatively small, as might generally

be supposed. We must therefore think of $(\mathcal{P}_{\text{sto}})$ as inherently a "large-scale" problem for which approximate methods of solution will be more appropriate than "exact" ones.

Nevertheless we would do well to keep in mind that the representation of u as a function from Ω to $\mathbb{R}^{k_0}\times\cdots\times\mathbb{R}^{k_T}$ tends to exaggerate the dimensionality of $(\mathcal{P}_{\text{sto}})$. The constraint that u_τ be $\mathcal{G}_\tau$-measurable means, as already noted, that u_τ can be identified uniquely with a certain function from $\mathcal{A}_\tau$ to $\mathbb{R}^{k_\tau}$. The dimension of the space of all functions from $\mathcal{A}_\tau$ to $\mathbb{R}^{k_\tau}$ is $a_\tau k_\tau$, where

$$a_k = |\mathcal{A}_k| \text{ (the number of atoms in } \mathcal{G}_k).$$

Thus the "true" dimensionality of $(\mathcal{P}_{\text{sto}})$, in the sense of the number of real-valued decision variables, is

$$k^* = a_0k_0 + a_1k_1 + \cdots + a_Tk_T. \tag{4.5}$$

By the same token, the "true" dimensionality of $(\mathcal{Q}_{\text{sto}})$, where the random variable $v=(v_1,\cdots,v_T,v_{T+1})$ must be optimized, is

$$l^* = a_1l_1 + \cdots + a_Tl_T + a_{T+1}l_{T+1}. \tag{4.6}$$

PROPOSITION 4.1. *Let*

$$\mathcal{U} = \{u=(u_0,u_1,\cdots,u_T) \mid u_\tau \text{ is } \mathcal{G}_\tau\text{-measurable with } u_\tau\in U_\tau\},$$
$$\mathcal{V} = \{v=(v_1,\cdots,v_T,v_{T+1}) \mid v_\tau \text{ is } \mathcal{G}_\tau\text{-measurable with } v_\tau\in V_\tau\},$$

and define $\mathcal{J}(u,v)=E\{J(u,v)\}$, *where* $J(u,v)$ *is the expression in Proposition* 3.2 (*regarded now as a random variable depending on the choice of the random variables* u *and* v). *Then* $\mathcal{U}$ *and* $\mathcal{V}$ *are polyhedral convex sets* (*nonempty*), *and* $\mathcal{T}$ *is a quadratic convex-concave function.*

Proof. By definition $\mathcal{U}$ is a subset of the space of all functions from Ω to $\mathbb{R}^{k_0}\times\cdots\times\mathbb{R}^{k_T}$ consisting of the functions u such that $u_{\omega\tau}\in U_{\omega\tau}$ for all ω and τ, and $U_{\omega\tau}$ is constant in ω with respect to each $\mathcal{G}_\tau$-atom $\alpha\in\mathcal{A}_\tau$. These conditions can be represented by a finite system of linear equations and inequalities, because Ω is finite and $U_{\omega\tau}$ is by assumption a convex polyhedron for each ω and τ. (Alternatively $\mathcal{U}$ can be viewed as a direct product of polyhedral convex sets $U_{\alpha\tau}$ indexed by $\alpha\in\mathcal{A}_\tau$ and $\tau=0,1,\cdots,T$, inasmuch as U_τ is $\mathcal{G}_\tau$-measurable.) Thus $\mathcal{U}$ is a convex polyhedron. Similarly $\mathcal{V}$ is a convex polyhedron. We have by definition

$$\mathcal{J}(u,v) = \sum_{\omega\in\Omega} \pi_\omega J(u_{\omega0},u_{\omega1},\cdots,u_{\omega T};v_{\omega1},\cdots,v_{\omega T},v_{\omega,T+1})$$

where the J term for each ω is a quadratic convex-concave function and the coefficients π_ω are nonnegative; therefore $\mathcal{J}$ is a quadratic convex-concave function. □

THEOREM 4.2. *The stochastic optimal control problems* $(\mathcal{P}_{\text{sto}})$ *and* $(\mathcal{Q}_{\text{sto}})$ *are the primal and dual problems of generalized linear-quadratic programming associated with the* $\mathcal{U}$, $\mathcal{V}$, *and* $\mathcal{J}$ *in Proposition* 4.1. *In particular, the assertions of Theorem* 2.1 *are valid for* $(\mathcal{P}_{\text{sto}})$ *and* $(\mathcal{Q}_{\text{sto}})$.

Proof. We must show that the supremum of $\mathcal{J}(u,v)$ over all $v\in\mathcal{V}$ is the function $f(u)$ in $(\mathcal{P}_{\text{sto}})$, and the infimum of $\mathcal{T}(u,v)$ over all $u\in\mathcal{U}$ is $g(u)$ in $(\mathcal{Q}_{\text{sto}})$. Starting with $J(u,v)$ in the form of (3.5) (which is obtained by using the right-hand expression in (3.3) for $[u,v]$) and taking the expectation, we get by (4.1) that

$$\begin{aligned}\mathcal{J}(u,v) = {} & \sum_{\tau=0}^{T} E\left\{p_\tau\cdot u_\tau + \frac{1}{2}u_\tau\cdot P_\tau u_\tau\right\} - \sum_{\tau=1}^{T+1} E\{c_\tau\cdot x_{\tau-1}\} \\ & + \sum_{\tau=1}^{T} E\left\{[q_\tau - E^{\mathcal{G}_\tau}\{C_\tau x_{\tau-1}\} - D_\tau u_\tau]\cdot v_\tau - \frac{1}{2}v_\tau\cdot Q_\tau v_\tau\right\} \\ & + E\{[q_{T+1} - E^{\mathcal{G}_{T+1}}\{C_{T+1}x_T\}]\cdot v_{T+1} - \tfrac{1}{2}v_{T+1}\cdot Q_{T+1}v_{T+1}\}.\end{aligned}$$

To maximize this over all $v \in \mathcal{V}$, we must maximize separately in each of the v_τ's subject to v_τ being a $\mathcal{G}_\tau$-measurable function with $v_\tau \in V_\tau$. Denote the random variable $q_\tau - E^{\mathcal{G}_\tau}\{c_\tau x_\tau\} - D_\tau u_\tau$ temporarily by r_τ for $\tau = 1, \cdots, T$ and $q_{T+1} - E^{\mathcal{G}_{T+1}}\{C_{T+1} x_T\}$ by r_{T+1}. Then each r_τ is $\mathcal{G}_\tau$-measurable and

$$\mathcal{J}(u, v) = \sum_{\tau=0}^{T} E\left\{p_\tau u_\tau + \frac{1}{2} u_\tau \cdot P_\tau u_\tau\right\} - \sum_{\tau=1}^{T+1} E\{c_\tau \cdot x_{\tau-1}\}$$
$$+ \sum_{\tau=0}^{T} \sup_{v_\tau \in \mathcal{V}_\tau} E\left\{r_\tau \cdot v_\tau - \frac{1}{2} v_\tau \cdot Q_\tau v_\tau\right\},$$

where $\mathcal{V}_\tau$ is the set of all $\mathcal{G}_\tau$-measurable v_τ with $v_\tau \in V_\tau$. Since $\mathcal{G}_\tau$-measurable functions can be indexed by $\alpha \in \mathcal{A}_\tau$ in place of $\omega \in \Omega$, as explained above, we can write

$$E\left\{r_\tau \cdot v_\tau - \frac{1}{2} v_\tau \cdot Q_\tau v_\tau\right\} = \sum_{\alpha \in \mathcal{A}_\tau} \pi_\alpha \left[r_{\alpha\tau} \cdot v_{\alpha\tau} - \frac{1}{2} v_{\alpha\tau} \cdot Q_{\alpha\tau} v_{\alpha\tau}\right],$$

where π_α is the probability of the atom α, i.e.,

$$\pi_\alpha = \sum_{\omega \in \alpha} \pi_\omega.$$

The supremum of this expression over all $v_\tau \in \mathcal{V}_\tau$ is

$$\sum_{\alpha \in \mathcal{A}_\tau} \pi_\alpha \sup_{v_{\alpha\tau} \in V_{\alpha\tau}} \left\{r_{\alpha\tau} \cdot v_{\alpha\tau} - \frac{1}{2} v_{\alpha\tau} \cdot Q_{\alpha\tau} v_{\alpha\tau}\right\} = \sum_{\alpha \in \mathcal{A}_\tau} \pi_\alpha \rho_{V_{\alpha\tau}, Q_{\alpha\tau}}(r_{\alpha\tau})$$
$$= E\{\rho_{V_\tau, Q_\tau}(r_\tau)\}.$$

Thus the supremum of $\mathcal{J}(u, v)$ over $v \in \mathcal{V}$ is

$$\sum_{\tau=0}^{T} E\left\{p_\tau \cdot u_\tau - \frac{1}{2} u_\tau \cdot P_\tau u_\tau\right\} - \sum_{\tau=1}^{T} E\{c_\tau \cdot x_{\tau-1}\} + \sum_{\tau=1}^{T+1} E\{\rho_{V_\tau, Q_\tau}(r_\tau)\},$$

which from choice of the r_τ's is the objective $f(u)$ in $(\mathcal{P}_{\text{sto}})$. The argument that the infimum of $\mathcal{J}(u, v)$ over $u \in \mathcal{U}$ is $g(v)$ in $(\mathcal{D}_{\text{sto}})$ follows the same lines. □

THEOREM 4.3. *For the $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{J}$ in Theorem 4.2 one has the following decomposability properties for separate minimization in u or maximization in v. The notation is used that*

$$\bar{v}_\tau = q_\tau - E^{\mathcal{G}_\tau}\{C_\tau \bar{x}_{\tau-1}\} - D_\tau \bar{u}_\tau \quad \text{for } \tau = 1, \cdots, T,$$
$$\bar{r}_{T+1} = q_{T+1} - E^{\mathcal{G}_{T+1}}\{C_{T+1} \bar{x}_T\},$$
$$\bar{s}_\tau = E^{\mathcal{G}_\tau}\{B_\tau^* \bar{y}_{\tau+1}\} + D_\tau^* \bar{v}_\tau - p_\tau \quad \text{for } \tau = 1, \cdots, T,$$
$$\bar{s}_0 = E^{\mathcal{G}_0}\{B_0^* \bar{y}_1\} - p_0,$$

where $\bar{u}$ and $\bar{v}$ are elements of $\mathcal{U}$ and $\mathcal{V}$, and $\bar{x}$ and $\bar{y}$ are the corresponding trajectories.

(a) $\tilde{u} \in \operatorname{argmin}_{u \in \mathcal{U}} \mathcal{J}(u, \bar{v})$ *if and only if*

$$\tilde{u}_{\alpha\tau} \in \partial \rho_{U_{\alpha\tau}, P_{\alpha\tau}}(\bar{s}_{\alpha\tau}) = \operatorname*{argmax}_{u_{\alpha\tau} \in U_{\alpha\tau}} \{\bar{s}_{\alpha\tau} \cdot u_{\alpha\tau} - \tfrac{1}{2} u_{\alpha\tau} \cdot P_{\alpha\tau} u_{\alpha\tau}\}$$

for $\tau = 0, 1, \cdots, T$ and all $\alpha \in \mathcal{A}_\tau$.

(b) $\tilde{v} \in \operatorname{argmax}_{v \in \mathcal{V}} \mathcal{J}(\bar{u}, v)$ *if and only if*

$$\tilde{v}_{\alpha\tau} \in \partial \rho_{V_{\alpha\tau}, P_{\alpha\tau}}(\bar{r}_{\alpha\tau}) = \operatorname*{argmax}_{v_{\alpha\tau} \in V_{\alpha\tau}} \{\bar{r}_{\alpha\tau} \cdot v_{\alpha\tau} - \tfrac{1}{2} v_{\alpha\tau} \cdot Q_{\alpha\tau} v_{\alpha\tau}\}$$

for $\tau = 1, \cdots, T, T+1$ and all $\alpha \in \mathcal{A}_\tau$.

Proof. This combines the argument of Theorem 4.2 with the conjugacy facts noted in the proof of Theorem 3.4. □

THEOREM 4.4. *Consider $\mathcal{G}_\tau$-measurable $\bar{u}$, $\bar{v}$, and the corresponding trajectories $\bar{x}$ and $\bar{y}$ determined by* (3.1) *and* (3.2). *Define the $\mathcal{G}_\tau$-measurable random variables*

$$\bar{p}_\tau = p_\tau - E^{\mathcal{G}_\tau}\{B_\tau^* \bar{y}_{\tau+1}\} \quad \text{for } \tau = 0, 1, \cdots, T,$$
$$\bar{q}_\tau = q_\tau - E^{\mathcal{G}_\tau}\{C_\tau \bar{x}_{\tau-1}\} \quad \text{for } \tau = 1, \cdots, T, T+1.$$

For each $\tau = 1, \cdots, T$ and $\alpha \in \mathcal{A}_\tau$ let $(\bar{\mathcal{P}}_{\alpha\tau})$ and $(\bar{\mathcal{Q}}_{\alpha\tau})$ denote the primal and dual problems of generalized linear-quadratic programming associated with

$$J_{\alpha\tau}(u_{\alpha\tau}, v_{\alpha\tau}) = \bar{p}_{\alpha\tau} \cdot u_{\alpha\tau} + \tfrac{1}{2} u_{\alpha\tau} \cdot P_{\alpha\tau} u_{\alpha\tau} + \bar{q}_{\alpha\tau} v_{\alpha\tau} - \tfrac{1}{2} v_{\alpha\tau} \cdot Q_{\alpha\tau} v_{\alpha\tau} - v_{\alpha}\tau \cdot D_{\alpha\tau} u_{\alpha\tau}$$

on $U_{\alpha\tau} \times V_{\alpha\tau}$, namely

$(\bar{\mathcal{P}}_{\alpha\tau})$ *minimize* $\bar{p}_{\alpha\tau} \cdot u_{\alpha\tau} + \tfrac{1}{2} u_{\alpha\tau} \cdot P_{\alpha\tau} u_{\alpha\tau} + \rho_{V_{\alpha\tau}, Q_{\alpha\tau}}(\bar{q}_{\alpha\tau} - D_{\alpha\tau} u_{\alpha\tau})$ *over* $u_{\alpha\tau} \in U_{\alpha\tau}$,

$(\bar{\mathcal{Q}}_{\alpha\tau})$ *maximize* $\bar{q}_{\alpha\tau} \cdot v_{\alpha\tau} - \tfrac{1}{2} v_{\alpha\tau} \cdot Q_{\alpha\tau} v_{\alpha\tau} - \rho_{U_{\alpha\tau}, P_{\alpha\tau}}(D_{\alpha\tau}^* v_{\alpha\tau} \bar{p}_{\alpha\tau})$ *over* $v_{\alpha\tau} \in V_{\alpha\tau}$,

and consider also the problems

$(\bar{\mathcal{P}}_{\alpha 0})$ *minimize* $\bar{p}_{\alpha 0} \cdot u_{\alpha 0} + \tfrac{1}{2} u_{\alpha 0} \cdot P_{\alpha 0} u_{\alpha 0}$ *over* $u_{\alpha 0} \in U_{\alpha 0}$

for $\alpha \in \mathcal{A}_0$, and

$(\bar{\mathcal{Q}}_{\alpha, T+1})$ *maximize* $\bar{q}_{\alpha, T+1} \cdot u_{\alpha, T+1} - \tfrac{1}{2} u_{\alpha, T+1} \cdot P_{\alpha, T+1}$ *over* $u_{\alpha, T+1} \in U_{\alpha, T+1}$

for $\alpha \in \mathcal{A}_{T+1}$.

Then a necessary and sufficient condition for $\bar{u}$ and $\bar{v}$ to be optimal solutions to the control problems $(\mathcal{P}_{\text{sto}})$ and $(\mathcal{Q}_{\text{sto}})$, respectively, is that $\bar{u}_{\alpha\tau}$ should be an optimal solution to the subproblem $(\bar{\mathcal{P}}_{\alpha\tau})$ for every $\alpha \in \mathcal{A}_\tau$ and $\tau = 0, 1, \cdots, T$, and $\bar{v}_{\alpha\tau}$ should be an optimal solution to the subproblem $(\bar{\mathcal{Q}}_{\alpha\tau})$ for every $\alpha \in \mathcal{A}_\tau$ and $\tau = 1, \cdots, T, T+1$.

Proof. The argument imitates the one for Theorem 3.5 but uses the relations in Theorem 4.3. □

REFERENCES

[1] R. T. ROCKAFELLAR, *Generalized linear-quadratic programming and optimal control*, SIAM J. Control Optim., 25 (1987), pp. 781–814.

[2] R. T. ROCKAFELLAR AND R. J-B WETS, *A dual solution procedure for quadratic stochastic programs with simple recourse*, in Numerical Methods, V. Pereyra and A. Reinoza, eds., Lecture Notes in Math. 1005, Springer-Verlag, New York, Berlin, 1983, pp. 252–265.

[3] ———, *A Lagrangian finite generation technique for solving linear-quadratic problems in stochastic programming*, Math. Programming Stud., 28 (1986), pp. 63–93.

[4] ———, *Linear-quadratic programming problems with stochastic penalties: the finite generation algorithm*, in Stochastic Optimization, V. Arkin, A. Shirayev, and R. Wets, eds., Lecture Notes in Control and Information Sci., IIASA Series No. 81, Springer-Verlag, New York, Berlin, 1986, pp. 454–560.

[5] A. KING, *An implementation of the finite generation method*, in Numerical Techniques for Stochastic Programming, Y. Ermoliev and R. J-B Wets, eds., Springer-Verlag, Heidelberg, 1987, pp. 295–311.

[6] R. W. COTTLE, *Symmetric dual quadratic programs*, Quart. Appl. Math., 21 (1963), pp. 237–243.

[7] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[8] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[9] R. GRINOLD, *Continuous programming part one: linear objectives*, J. Math. Anal. Appl., 37 (1972), pp. 130–141.

[10] D. BERTSEKAS, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.

[11] M. EISNER AND P. OLSEN, *Duality for stochastic programming interpreted as* L.P. *in Lp-space*, SIAM J. Appl. Math., 28 (1975), pp. 779–792.

[12] R. T. ROCKAFELLAR AND R. WETS, *The optimal recourse problem in discrete time: L^1-multipliers for inequality constraints*, SIAM J. Control Optim., 16 (1978), pp. 16–36.

[13] D. BERTSEKAS AND S. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 823-840, July 1990

004

SINGULAR OPTIMAL CONTROL PROBLEMS: ON THE NECESSARY CONDITIONS OF OPTIMALITY*

F. LAMNABHI-LAGARRIGUE† AND G. STEFANI‡

Abstract. The purpose of this paper is to give higher-order necessary conditions for the optimality of a totally singular arc in the case of nonlinear systems not necessarily linear in the control variables. These conditions, stated in Theorems 2 and 3, are all expressed in terms of the derivative of a particular function along suitable Lie brackets involving vector fields associated with the control process. These conditions contain most of the early results and in addition some of them are new, especially those of third order. To prove these results two different methods are used, namely, Sussmann's techniques arising in local controllability theory *and* Volterra series expansion.

Key words. singular optimal control, totally singular arcs, second- and third-order necessary conditions, local controllability, Volterra series

Introduction. The term singular is used in optimal control problems in which the Pontryagin maximum principle [17] does not furnish an explicit relationship between the control and the state and costate variables (for introductory material see Bell and Jacobson [2]). Many practical applications, namely, in rocket and air vehicle flight, exhibit solutions that include singular arcs (see, for instance, Kelley, Kopp, and Moyer [11] or Vinh [22]). Singular optimal control problems may also be found in heat transfer control problems [16]. These problems have been an active research area for two decades. Many techniques have been used (see, for instance, Agracev [1], Brockett [3], Gabasov and Kirillova [6], Goh [7], Gorokhovik [8], Jacobson and Speyer [9], Kazemi-Dehkordi [10], Kelley, Kopp, and Moyer [11], Knobloch [12], Krener [13], Lamnabhi-Lagarrigue [14], and Wagner [23] and the bibliography therein). However, there is a need to unify these results and to go further in the analysis, i.e., to obtain third- and higher-order necessary conditions. There exist a few third-order conditions for problems with terminal constraints, as in Krener's work [13] and in a paper by Wagner [23].

We consider here nonlinear systems not necessarily linear in the control variables, and we do not limit ourselves to scalar controls. However, we do not consider terminal state constraints; necessary conditions for optimality for fixed-endpoint problems are considered, for instance, in Gorokhovik [8], Knobloch [12], and Krener [13]. We also limit our study to totally singular arcs in the C^∞ case. The material contained in this paper will be used in a future publication for investigating both fixed-endpoint problems and partially singular arcs.

The purpose of this paper is to state two theorems containing second- and third-order necessary conditions for optimality. Theorem 2 appears as the first step of Theorem 3 ($s = 0$). It contains necessary conditions that can be derived using only the fact that the reference trajectory is singular. Theorem 3 allows us to consider inductively more degenerated situations, giving a sequence of necessary conditions. Each time a nonnegative quadratic form is zero, some other directional derivatives are zero and a new nonnegative quadratic form is obtained.

* Received by the editors November 30, 1987; accepted for publication (in revised form) March 21, 1989. This work was completed while the first author was visiting the Department of Electrical Engineering and the Department of Mathematics, Arizona State University at Tempe, Arizona.

† Laboratoire des Signaux et Systèmes, ESE, Plateau du Moulon, 91190 Gif-sur-Yvette, France.

‡ Dipartamento di Matematica e Applicasioni, Via Mezzocannone 8, 80100, Napoli, Italia.

This work follows from two earlier papers by Lamnabhi-Lagarrigue [14], [15]. The results of the first paper are generalized here and more details are included. The second paper and the present one constitute, in our opinion, a new setting for singular optimal control problems, or more precisely, for higher-order optimality conditions. It is interesting to note that the starting point of this approach is a recent result of Sussmann studying local controllability of nonlinear systems [21]. This result is a key tool in that it provides a suitable control variation in the proof of parts (i) and (ii) of Theorem 2 and part (ii) of Theorem 3. This again demonstrates a similarity between the problems of finding sufficient conditions for local controllability and of finding necessary conditions for optimality. The remaining results of the paper are proved by combining Volterra series expansions, special control variation, and multiple integral identities. This combination of material has also been used recently by Stefani [18], [19] for deriving a sufficient condition for extremality and to extend part of the results to the case in which the system or the extremal trajectory are not smooth.

1. Statement of the problem. Let us consider the control system

$$\Sigma\begin{cases}\dot{x}(t)=f(x(t), u(t)),\\ x(0)=x^0,\end{cases}$$

$x\in\mathbf{R}^n$, where the set of admissible controls $\mathcal{U}$ is the set of the integrable functions $t\to u(t)$ that take values in some bounded open set $U\subset\mathbf{R}^m$ and where f: $\mathbf{R}^n\times\mathbf{R}^m\to\mathbf{R}^n$ is a C^∞ mapping. In particular $f(\cdot, v)$ is a C^∞ vector field for each $v\in U$. Let $h:\mathbf{R}^n\to\mathbf{R}$ be a smooth function and let $t\to x(x^0, u, t)$ be the solution relative to the control u of Σ.

The control problem under consideration here can be stated as follows. Let $t\to\bar{u}(t)$ be a given C^∞ control, $\bar{u}(t)\in U$, and let T be a fixed terminal time. Find necessary conditions such that

$$h(x(x^0, \bar{u}, T))=\min_{u\in\mathcal{U}} h(x(x^0, u, T));$$

$\bar{u}(t)$ is called the reference control.

Adding a new coordinate to Σ, say x_0, and an equation $\dot{x}_0=1$, it is not difficult to see that we can choose $\bar{u}(t)=0$, $t\in[0, T]$ to be our reference control. We will denote by $\gamma(t)$ the associated reference trajectory:

$$\gamma(t)=x(x^0, \bar{u}, t)=x(x^0, u, t), \qquad t\in[0, T].$$

Moreover, we can assume that $\mathcal{U}$ is the set of integrable functions from $[0, T]$ to the set $\Omega=\{(v_1, \cdots, v_m)\in\mathbf{R}^m: |v_i|<1\}$. Therefore the stated optimal control problem takes the following form. Find necessary conditions such that

$$h(\gamma(T))=\min_{u\in\mathcal{U}} h(x(x^0, u, T)). \tag{1}$$

If we introduce the vector field $f_0:\mathbf{R}^n\to\mathbf{R}^n$

$$f_0(x)=f(x, 0)$$

we can write (1) as

$$h(\exp Tf_0\cdot x^0)=\min_{u\in\mathcal{U}} h(x(x^0, u, T)) \tag{1'}$$

where $(x, t)\to\exp tf_0\cdot x$ is the local flow of the vector field f_0.

2. First-order necessary conditions. The Hamiltonian and the adjoint system associated with Σ are, respectively,

$$\mathbf{H}(x, u, \lambda) = \langle \lambda, f(x, u) \rangle$$

and

$$\dot{\lambda}(t) = -\frac{\partial \mathbf{H}}{\partial x}(\gamma(t), 0, \lambda(t)) \equiv \lambda(t) \frac{\partial f_0}{\partial x}(\gamma(t)), \qquad \lambda(T) = dh(\gamma(T)). \tag{2}$$

For the problem stated in § 1, the Maximum Principle [17] gives the following necessary conditions for optimality.

THEOREM 1. *If the reference control $\bar{u}(t) = 0$ is minimal on $[0, T]$, i.e., it satisfies* (1), *then there exists an adjoint vector $\lambda(t)$, solution of the adjoint equation* (2) *such that*

$$\frac{\partial \mathbf{H}}{\partial u}(\gamma(t), 0, \lambda(t)) = 0 \quad \text{for } t \in [0, T]$$

and

$$\frac{\partial^2 \mathbf{H}}{\partial u^2}(\gamma(t), 0, \lambda(t)) = \left(\left(\frac{\partial^2 \mathbf{H}}{\partial u_i \, \partial u_j}(\gamma(t), 0, \lambda(t)) \right) \right)_{1 \leqq i \leqq m, 1 \leqq j \leqq m} \tag{3}$$

is a nonnegative matrix for $t \in [0, T]$.

Condition (3) is usually called the Legendre-Clebsch condition.

Before stating the main results and then introducing some technical tools for proving them, let us recall some known results.

Let $w_0 : [0, T] \times \mathbf{R}^n \to R$ be defined as follows:

$$w_0(t, x) = h(\exp(T - t) f_0 \cdot x) \quad \text{for } t \in [0, T].$$

In particular, note that

$$w_0(t, \gamma(t)) = h(\gamma(T)). \tag{4}$$

LEMMA 1. *The map $\lambda : [0, T] \to (\mathbf{R}^n)^*$ given by $\lambda(t) = \partial w_0 / \partial x(t, \gamma(t))$ is the solution of the adjoint equation*

$$-\dot{\lambda}(t) = \lambda(t) \frac{\partial f_0}{\partial x}(\gamma(t)) \quad \text{and} \quad \lambda(T) = dh(\gamma(t)).$$

Let $g \cdot w_0(t, x)$ denote $\langle \partial w_0 / \partial x(t, x), g(x) \rangle$ where $g(x)$ is a vector field.

The following result can also be proved.

LEMMA 2. *If $g(x)$ is a vector field such that*

$$g \cdot w_0(t, \gamma(t)) = 0 \quad \text{for } t \in [0, T],$$

then

$$\mathrm{ad}_{f_0}^k g \cdot w_0(t, \gamma(t)) = 0 \quad \text{for } t \in [0, T] \text{ and } k \geqq 1,$$

where $\mathrm{ad}_{f_0}^k g$ is recursively defined by $\mathrm{ad}_{f_0}^0 g = g$ and $\mathrm{ad}_{f_0}^k g = [f_0, \mathrm{ad}_{f_0}^{k-1} g]$.

Now let $f_i(x)$ and $f_{ij}(x)$ be, respectively, the vector fields $\partial f / \partial u_i(x, 0)$ and $\partial^2 f / \partial u_i \, \partial u_j(x, 0)$. From Theorem 1 and Lemmas 1 and 2 we obtain Corollary 1.

COROLLARY 1. *If the reference control $\bar{u}(t) = 0$ is minimal on $[0, T]$, then*

$$\mathrm{ad}_{f_0}^k f_i \cdot w_0(t, \gamma(t)) = 0 \quad \text{for } t \in [0, T]$$

and the matrix

$$((f_{ij} \cdot w_0(t, \gamma(t))))_{1 \leqq i \leqq m, 1 \leqq j \leqq m} \tag{3'}$$

is a nonnegative matrix for $t \in [0, T]$.

Theorem 1 (or Corollary 1) is also called first-order necessary conditions. The purpose of the following sections is to state and prove analogous results for singular optimal control problems by deriving so-called second- and third-order necessary conditions. As in Corollary 1 these conditions are all expressed in terms of the derivative of the function $w_0(t, .)$ along suitable Lie brackets involving vector fields associated with the control process, such as, for instance, f_i, f_{ij}.

3. Singular optimal control—statement of the main results. Let $\gamma(t)$ be an extremal reference trajectory on $[0, T]$, that is,

$$\frac{\partial \mathbf{H}}{\partial u}(\gamma(t), 0, \lambda(t)) = 0 \quad \text{for } t \in [0, T].$$

DEFINITION 1. The extremal reference trajectory $\gamma(t)$ is said to be *totally singular* on $[a, b] \subset [0, T]$ if and only if

$$\frac{\partial^2 \mathbf{H}}{\partial u^2}(\gamma(t), 0, \lambda(t)) = 0 \quad \text{for } t \in [a, b]$$

or equivalently if

$$f_{ij} w_0(t, \gamma(t)) = 0 \quad \text{for } t \in [a, b] \text{ and } i, j \in \{1, \cdots, m\}.$$

DEFINITION 2. The extremal reference trajectory $\gamma(t)$ is said to be *partially singular* on $[a, b] \subset [0, T]$ if and only if

$$\det\left(\frac{\partial^2 \mathbf{H}}{\partial u^2}(\gamma(t), 0, \lambda(t))\right) = 0 \quad \text{for } t \in [a, b]$$

and

$$\frac{\partial^2 \mathbf{H}}{\partial u^2}(\gamma(t), 0, \lambda(t)) \not\equiv 0 \quad \text{for } t \in [a, b].$$

In the following we are concerned only with totally singular arcs and we assume that $[a, b]$ is the whole interval $[0, T]$. The reference trajectory $\gamma(t)$ will be called a singular trajectory. Let us first introduce some notation.

For each $u \in \mathcal{U}$, and each $p > 1$,

$$\|u\|_p := \int_0^T \left(\sum_{i=1}^m |u_i(t)|^p \, dt \right)^{1/p}.$$

Moreover, for each multi-index $\nu = (\nu_1, \nu_2, \cdots, \nu_m)$ with $\nu_i \geqq 0$, we define the length of ν by $|\nu| := \sum_{i=1}^m \nu_i$, $\nu! := \nu_1! \nu_2! \cdots \nu_m!$, $v^\nu := v_1^{\nu_1} . v_2^{\nu_2} . \cdots . v_m^{\nu_m}$ for each $v = (v_1, \cdots, v_m) \in \mathbf{R}^m$ and $f_\nu(x)$ (or simply f_ν) will denote the vector field

$$\frac{\partial^{|\nu|}}{(\partial u_1)^{\nu_1} \cdots (du_m)^{\nu_m}} f(x, 0). \tag{5}$$

Using this notation, note that the vector fields $f_0(x)$, $f_i(x)$, and $f_{ij}(x)$ defined above stand, respectively, for $f_{(0,\cdots,0)}(x)$, $f_{(0,\cdots,\underset{i}{1},0)}(x)$, and $f_{(0,\cdots,\underset{i}{1},0,\cdots,\underset{j}{1},0,\cdots)}(x)$.

To state the main results we introduce some additional notation. Let $A^{(r)}$ be the nilpotent algebra of step r of polynomials in the noncommutative indeterminates

$$\{X_\nu, |\nu| = 0, \cdots, \ell\}.$$

The nilpotent Lie algebra of step r, $L^{(r)}$, in the same set of indeterminates can be viewed as a subset of $A^{(r)}$. In the ideal $\mathcal{J}^{(r)}$ generated in $L^{(r)}$ by $\{X_\nu, |\nu| = 1, \cdots, \ell\}$, we define weights for the brackets by

$$\begin{aligned} &\omega_i(X_\nu) = \nu_i \\ &\omega_i(X_0) = 0 \end{aligned} \quad \text{if } \nu = (\nu_1, \cdots, \nu_m\}, \quad i = 1, \cdots, m$$

and

$$\omega_i([Z_1, Z_2]) = \omega_i(Z_1) + \omega_i(Z_2) \quad \text{for } Z_1, Z_2 \in L^{(r)}.$$

Finally, $\omega(Y) := \sum_{i=1}^m \omega_i(Y)$ and $\|Y\|$ will denote the length of Y, i.e., the number of brackets involved in Y, $Y \in \mathcal{J}^{(r)}$. For each $\Lambda \in L^{(r)}$ we obtain a vector field ev Λ by substituting f_ν for X_ν. If $\phi = \text{ev}\,\Lambda$, $\omega(\Lambda)$ will indicate in which Volterra kernel the vector field ϕ appears and therefore it will be directly linked to the "order" of the necessary condition. $\|\Lambda\|$ is the number of vector fields involved.

We can now state the main results.

THEOREM 2. *Let $\gamma(t)$ be an extremal singular trajectory on $[0, T]$. If $\gamma(t)$ is minimal, i.e., satisfies* (1), *then*

(i) $\text{ad}_{f_0}^k [f_i, f_j] w_0(t, \gamma(t)) = 0$ *for* $i, j \in \{1, \cdots, m\}$, $k \geqq 0$ *and* $t \in [0, T]$.

(ii) $\text{ad}_{f_0}^k f_\nu \cdot w_0(t, \gamma(t)) = 0$ *for* $|\nu| = 3$, $k \geqq 0$ *and* $t \in [0, T]$.

(iii) *The matrix*

$$((([\text{ad}_{f_0} f_i, f_j] \cdot w_0(t, \gamma(t))))_{1 \leqq j \leqq m, 1 \leqq j \leqq m}$$

is a symmetric nonnegative matrix for $t \in [0, T]$.

Remark 1. Each of the relations in Theorem 2 may be interpreted in terms of the Hamiltonian $\mathbf{H}$. For instance, condition (i) for $k = 0$ is equivalent to

$$\frac{\partial^2 \mathbf{H}}{\partial x\, \partial u_j}(\gamma(t), 0, \lambda(t)) \frac{\partial f}{\partial u_i}(\gamma(t), 0) - \frac{\partial^2 \mathbf{H}}{\partial x\, \partial u_i}(\gamma(t), 0, \lambda(t)) \frac{\partial f}{\partial u_j}(\gamma(t), 0) = 0.$$

This condition can be found, for instance, in Gabasov and Kirillova [6, p. 144]. Similarly, condition (ii) means

$$\frac{d^k}{dt^k} \frac{\partial^3 \mathbf{H}}{\partial u_i\, \partial u_j\, \partial u_l}(\gamma(t), 0, \lambda(t)) = 0 \quad \text{for } i, j, l \in \{1, \cdots, m\}, k \geqq 0 \text{ and } t \in [0, T].$$

To our knowledge this condition is new.

Finally, condition (iii) can be shown easily to be equivalent to

$$\sum_{i,j=1}^m \frac{\partial}{\partial u_i} \frac{d^2}{dt^2} \frac{\partial \mathbf{H}}{\partial u_j}(\gamma(t), 0, \lambda(t)) \eta_i \eta_j \geqq 0$$

for $\eta_i \eta_j \geqq 0$ and $t \in [0, T]$. This is the well-known so-called *generalized Legendre-Clebsch* condition that has been derived by several authors. For fixed-endpoint singular optimal control problems, this condition is also derived in Knobloch [12] and Krener [13] without any assumption of normality.

Remark 2. From Lemma 2 and Definition 1 we know that if $\gamma(t)$ is a singular arc, then $\text{ad}_{f_0}^k f_{ij} \cdot w_0(t, \gamma(t)) = 0$ for $t \in [0, T]$ and $i, j \in \{1, \cdots, m\}$. However, note that this condition is definitely different from (i) of Theorem 2.

THEOREM 3. *Let $\gamma(t)$ be an optimal singular trajectory and assume that there exists $s \geqq 1$ such that for $t \in [0, T]$ and $i, j \in \{1, \cdots, m\}$,*

$$[\text{ad}_{f_0}^{k+1} f_i, \text{ad}_{f_0}^k f_j] \cdot w_0(t, \gamma(t)) = 0 \tag{6}$$

for $k = 0, \cdots, s-1$.

If $\gamma(t)$ is minimal, then

(i) $[\mathrm{ad}_{f_0}^{k_1} f_i, \mathrm{ad}_{f_0}^{k_2} f_j] \cdot w_0(t, \gamma(t)) = 0$ *for* $t \in [0, T]$, $i, j \in \{1, \cdots, m\}$ and $k_1, k_2 \geqq 0$ *such that* $k_1 + k_2 = 0, \cdots, 2s$.

(ii) $\mathrm{ad}_{f_0}^{k} \phi \cdot w_0(t, \gamma(t)) = 0$ *for* $t \in [0, T]$, $k \geqq 0$ *and* $\phi = \mathrm{ev}\, \Lambda$ *with* $\Lambda \in \mathcal{P}^{(s+1)}$ *such that* $\omega(\Lambda) = 3$.

(iii) *The matrix*

$$([\mathrm{ad}_{f_0}^{s+1} f_i, \mathrm{ad}_{f_0}^{s} f_j] \cdot w_0(t, \gamma(t)))_{1 \leqq i \leqq m, 1 \leqq j \leqq m}$$

is a symmetric nonnegative matrix for $t \in [0, T]$.

Remark 3. Since Theorems 2 and 3 give necessary conditions, they are also valid if the set of constraints for the control is an unbounded neighbourhood of zero in $\mathbf{R}^m$.

Remark 4. Let us give some examples to demonstrate Theorem 3 more explicitly. If

$$[\mathrm{ad}_{f_0} f_i, f_j] \cdot w_0(t, \gamma(t)) = 0 \quad \text{for } t \in [0, T], \quad i, j \in \{1, \cdots, m\}, \tag{7}$$

i.e., condition (ii) of Theorem 2 is trivially satisfied, then (i) and (ii) of Theorem 2 hold and moreover the following conditions are also satisfied:

(α) $[\mathrm{ad}_{f_0}^{k_1} f_i, \mathrm{ad}_{f_0}^{k_2} f_j] \cdot w_0(t, \gamma(t)) = 0$ for $t \in [0, T]$, $i, j \in \{1, \cdots, m\}$ and $k_1 + k_2 \leqq 2$.

(β) $\mathrm{ad}_{f_0}^{k} [f_{ij}, f_l] \cdot w_0(t, \gamma(t)) = 0$ for $t \in [0, T]$, $k \geqq 0$ and $i, j, l \in \{1, \cdots, m\}$.

(γ) The matrix

$$([\mathrm{ad}_{f_0}^{2} f_i, \mathrm{ad}_{f_0} f_j] w_0(t, \gamma(t)))_{1 \leqq i \leqq m, 1 \leqq j \leqq m} \tag{8}$$

is a symmetric nonnegative matrix for $t \in [0, T]$.

Similarly, assume now that in addition to (7) we have $[\mathrm{ad}_{f_0}^{2} f_i, f_j] \cdot w_0(t, \gamma(t)) = 0$ for $t \in [0, T]$ and $i, j \in \{1, \cdots, m\}$, i.e., condition (8) is trivially satisfied. Then from Theorem 3 the following conditions are satisfied:

(α') $[\mathrm{ad}_{f_0}^{k_1} f_i, \mathrm{ad}_{f_0}^{k_2} f_j] \cdot w_0(t, \gamma(t)) = 0$ for $t \in [0, T]$, $i, j \in \{1, \cdots, m\}$ and $k_1 + k_2 \leqq 4$.

(β') $\mathrm{ad}_{f_0}^{k} [f_i, [f_j, f_l]] \cdot w_0(t, \gamma(t)) = 0$,
$\mathrm{ad}_{f_0}^{k} [f_{ij}, [f_0, f_l]] \cdot w_0(t, \gamma(t)) = 0$,
$\mathrm{ad}_{f_0}^{k} [[f_0, f_{ij}], f_l] \cdot w_0(t, \gamma(t)) = 0$,
for $t \in [0, T]$, $k \geqq 0$ and $i, j, l \in \{1, \cdots, m\}$.

(γ') The matrix

$$([\mathrm{ad}_{f_0}^{3} f_i, \mathrm{ad}_{f_0}^{2} f_j] \cdot w_0(t, \gamma(t)))_{1 \leqq i \leqq m, 1 \leqq j \leqq m}$$

is a symmetric nonnegative matrix.

Remark 5. As in Remark 1, we may interpret these conditions in terms of the Hamiltonian $\mathbf{H}$. For instance, condition (β) can be formulated as follows:

$$\frac{\partial^2 \mathbf{H}}{\partial x\, \partial u_l}(\gamma(t), 0, \lambda(t)) \frac{\partial^2 f}{\partial u_i\, \partial u_j}(\gamma(t), 0) = \frac{\partial^3 \mathbf{H}}{\partial x\, \partial u_i\, \partial u_j}(\gamma(t), 0, \lambda(t)) \frac{\partial f}{\partial u_l}(\gamma(t), 0)$$

for all $t \in [0, T]$ and $i, j, l \in \{1, \cdots, m\}$.

Note also that part (i) of Theorem 3 is already contained in Knobloch [12, § 22/23].

4. Some preliminary results. The purpose of this section is to show how it is possible to approximate a solution of the system Σ by solutions of other systems that have useful properties enabling us to prove the main results (see also Crouch [5]).

Note first that for each $\nu=(\nu_1, \nu_2, \cdots, \nu_m)$, $x \to (\partial^{|\nu|}/(\partial u_1)^{\nu_1}\cdots(\partial u_m)^{\nu_m})f(x,0)=f_\nu(x)$ is a C^∞ vector field and that for each compact *nbh* $K \subset M$ of x^0, there exists a constant $\rho>0$ such that

$$\left\| f(x,v)-f_0(x)-\sum_{|\nu|=1}^{l} f_\nu(x)\frac{v^\nu}{\nu!}\right\| \leqq \rho\left(\sum_{i=1}^{m}|v_i|\right)^{l+1}$$

for each $x\in K$ and $v\in\Omega$. We have the following standard result.

LEMMA 3. *For each $H>0$ there exists a $\sigma>0$ such that if $\|u\|_1 \leqq \sigma$ then $x(x^0, u, .)$ is defined on $[0, T]$ and $\|x(x^0, u, t)-\gamma(t)\| \leqq H$ for $t\in[0, T]$.*

Let us now associate to Σ the following system Σ^l:

$$\Sigma^l\begin{cases}\dot z(t)=f_0(z(t))+\sum_{|\nu|=1}^{l} f_\nu(z(t))\dfrac{u^\nu(t)}{\nu!},\\ z(0)=x^0,\end{cases}$$

where $f(x, v)$ is approximated by its Taylor expansion up to order l. The next result compares a solution of the system Σ and a solution of the system Σ^l with the same initial condition x^0.

LEMMA 4. *There exist σ and $H>0$ such that, if $\|u\|_i \leqq \sigma$, $i=1, \cdots, l$, then*

$$\|x(x^0, u, t)-z(x^0, u, t)\| \leqq H\|u\|_{l+1}^{l+1} \quad \textit{for } t\in[0, T],$$

where $x(x^0, u, t)$ is the solution of Σ relative to the control u, and $z(x^0, u, t)$ is the solution of Σ^l relative to the same control u, both systems initialized at x^0.

Proof. We first remark that if $\|u\|_1$ is sufficiently small, then $x(x^0, u, .)$ and $z(x^0, u, .)$ are defined on $[0, T]$ and belong to a suitable compact neighbourhood K of x^0. Moreover,

$$\|x(x^0,u,t)-z(x^0,u,t)\| \leqq \int_0^t \left\{ \left\| f(x(x^0,u,s),u(s))-\sum_{|\nu|=0}^{l} f_\nu(x(x^0,u,s))\frac{u^\nu(s)}{\nu!}\right\| + \sum_{|\nu|=1}^{l}\|f_\nu(x(x^0,u,s))-f_\nu(z(x^0,u,s))\|\frac{|u^\nu(s)|}{\nu!}\right\} ds.$$

But $|u^\nu(s)|<1$ almost everywhere on $[0, T]$, hence

$$\|x(x^0,u,t)-z(x^0,u,t)\| \leqq \rho\int_0^t \|u(s)\|^{l+1}\,ds+\sum_{|\nu|=0}^{l}\int_0^t H_1\|x(x^0,u,s)-z(x^0,u,s)\|\,ds.$$

The statement follows using the Gronwall inequality. □

Let us now approximate the solutions of Σ^l by the solutions of a system Σ_r^l defined on a nilpotent Lie group. To be more precise let $A^{(r)}$ and $L^{(r)}$ be defined as in § 3, and let $G^{(r)}=\{\exp Z=\sum_{i=0}^{r} Z^i/i!,\ Z\in L^{(r)}\}$ be the nilpotent Lie group associated with $L^{(r)}$. Each $Z\in L^{(r)}$ can be identified with a vector field on $A^{(r)}$ as follows:

$$S \to SZ \quad \text{(the product of } S \text{ and } Z \text{ in } A^{(r)}).$$

It is well known that the system Σ_r^l defined on $A^{(r)}$

$$\Sigma_r^l\begin{cases}\dot S=S\left(X_0+\sum_{|\nu|=1}^{l} u^\nu(t)X_\nu\right),\\ S(0)=I\end{cases}$$

evolves on $G^{(r)}$ (see [21]).

In the following lemma we state that the set of reachable points by Σ_r^l has nonempty interior in $G^{(r)}$ (see [20]).

LEMMA 5. *The Lie algebra generated by*

$$\left\{ X_0 + \sum_{|\nu|=1}^{l} v^\nu X_\nu,\ v \in \Omega \right\}$$

is identical to $L^{(r)}$

Proof. It is sufficient to prove that

$$X_\mu \in \operatorname{span} \left\{ X_0 + \sum_{|\nu|=1}^{l} v^\nu X_\nu,\ v \in \Omega \right\}$$

for each μ such that $|\mu| = 1, \cdots, l$. Assume that this is not true; then there exists a linear form ω such that

$$\langle \omega, X_\mu \rangle \neq 0 \tag{9}$$

and

$$\left\langle \omega, X_0 + \sum_{|\nu|=1}^{l} v^\nu X_\nu \right\rangle = 0, \qquad v \in \Omega.$$

From this identity we obtain

$$\left\langle \omega, \frac{\partial^\mu}{\partial v^\mu} \left(\sum_{|\nu|=1}^{l} v^\nu X_\nu \right)_{|v=0} \right\rangle = 0.$$

Therefore $\langle \omega, X_\mu \rangle = 0$, a contradiction. □

We can also prove that the ideal generated by

$$\left\{ \sum_{|\nu|=1}^{l} v^\nu X_\nu,\ v \in \Omega \right\} \quad \text{in } L^{(r)}$$

is identical to the ideal $\mathcal{I}^{(r)}$.

As we did before for each $\Lambda \in L^{(r)}$, we obtain for each $S \in A^{(r)}$ a differential operator by substituting f_ν for X_ν. Following Sussmann [21] we denote this differential operator by ev S. $\operatorname{ev}_y S$ will denote ev S evaluated at $y \in M$. For example, for each C^∞ function ϕ

$$\operatorname{ev}_y \exp tX_0 \cdot \phi = \sum_{i=0}^{r} \frac{t^i}{i!} f_0^i \cdot \phi(y).$$

Moreover, if $x = (x_1, \cdots, x_n)$ are the coordinate functions of $\mathbf{R}^n$, then

$$\operatorname{ev}_y S \cdot x = (\operatorname{ev}_y S \cdot x_1, \cdots, \operatorname{ev}_y S \cdot x_n).$$

Let $S(I, u, t)$ denote the solution of Σ_r^l relative to the control u and with initial condition the identity of $G^{(r)}$. Let $\Gamma(S^0, -t)$ denote the solution at times $-t$ of

$$\dot{S} = SX_0, \qquad S(0) = S^0.$$

We easily obtain $\Gamma(S^0, -t) = S^0 \Gamma(I, -t) = S^0 \exp(-tX_0)$.

LEMMA 6. Sussmann [21]. *For t sufficiently small and $y \in K$, a compact of $\mathbf{R}^n$, there exists $H > 0$ such that*

$$\| \operatorname{ev}_y \Gamma(S(I, u, t), -t) \cdot x - \exp(-tf_0) \cdot z(y, u, t) \| \leqq Ht^{r+1}.$$

Then from Lemma 4 we obtain

$$\|\mathrm{ev}_y\,\Gamma(S(I,u,t),-t)\cdot x-\exp(-tf_0)\cdot x(y,u,t)\|\leqq H(t^{r+1}+\|u\|_{l+1}^{l+1}) \tag{10}$$

for $y\in\gamma([0,T])$.

Let $N_r=\dim\mathcal{J}^{(r)}$ and let $X=\{Y_1,\cdots,Y_{N_r}\}$ be a basis of $\mathcal{J}^{(r)}$ whose elements are the $X_\nu s$ and their brackets. Each element of $G^{(r)}$ is the exponential of an element in $L^{(r)}$. Therefore we may write $\Gamma(\mathrm{S(I,u,t)},-\mathrm{t})$ as

$$\exp\sum_{j=1,\cdots,N_r}p(j,u,t)\,Y_j \tag{11}$$

where $t\to p(j,u,t)$ are suitable absolutely continuous functions.

We will say that $Y\in\mathcal{J}^{(r)}$ is even if $\omega_i(Y)$ is even for *each* $i=1,\cdots,m$. (The ω_i's are defined in § 3.)

If $Y_j\in X$ we set $\omega(Y_j)=\beta_j$ and $\|Y_j\|=\alpha_j$.

Using Sussmann's arguments [21], we can state the following result that is at the basis of the construction of most of the variations used later.

THEOREM 4. *There exist* $\bar t>0$, $\bar u\in\mathcal{U}$, *a ball* $B_\rho\subset\mathbf{R}^{N_r}$, *and a map* $\xi:B_\rho\to\mathcal{U}$ *such that*

(i) $\xi(0)=\bar u\in\mathcal{U}$.

(ii) $\Gamma(S(I,\bar u,\bar t),-\bar t)=S(I,\bar u,\bar t)\exp(-\bar tX_0)=\exp\sum_{i=1,\cdots,N_r}q(i,\bar u,\bar t)\,Y_i$ *where* $q(i,\bar u,\bar t)$ *are suitable coefficients that are zero if* Y_i *is not even.*

(iii) *If* $c\in B_\rho$, *then*

$$\Gamma(S(I,\xi(c),\bar t),-\bar t)=\exp\left\{\sum_{i=1,\cdots,N_r}q(i,\bar u,\bar t)\,Y_i+c_iY_i\right\}.$$

Remark 6. Looking at Sussmann's proof, we note that the function ξ is continuous with respect to each norm $\|\ \|_p$, $p\neq\infty$. However, we do not need this property. Indeed, we use (iii) of Theorem 4 for a fixed $c\in B_\rho$ and we need only the existence of ξ and the property of $p(j,u,t)$ stated in the following lemma.

LEMMA 7. *Let* $\bar t$ *be any positive real number, let* $u:[0,\bar t]\to\mathbf{R}^m$ *be a control, and let* $u_\varepsilon:[0,\varepsilon^a\bar t]\to\mathbf{R}^m$ *defined by* $u_\varepsilon(\tau)=\varepsilon^bu(\tau/\varepsilon^a)$ *where a and b are some positive parameters. Then if* $u_{\varepsilon k}$ *denotes the kth component of* u_ε

(i)
$$\int_0^{\varepsilon^a\bar t}\tau^{k_2}u_{\varepsilon j}(\tau)\left(\int_0^\tau\sigma^{k_1}u_{\varepsilon i}(\sigma)\,d\sigma\right)d\tau$$
$$=\varepsilon^{a(k_1+k_2+2)+2b}\int_0^{\bar t}\tau^{k_2}u_j(\tau)\left(\int_0^\tau\sigma^{k_1}u_i(\sigma)\,d\sigma\right)d\tau$$

and

(ii)
$$\Gamma(S(I,u_\varepsilon,\varepsilon^at),-\varepsilon^at)=\exp\sum_{j=1,\cdots,N_r}p(j,u_\varepsilon,\varepsilon^at)\,Y_j$$
$$=\exp\sum_{j=1,\cdots,N_r}\varepsilon^{a\alpha_j+b\beta_j}p(j,u,t)\,Y_j.$$

Proof. The definition of u_ε gives

$$\int_0^{\varepsilon^a\bar t}\tau^{k_2}\varepsilon^bu_j\left(\frac{\tau}{\varepsilon^a}\right)\left(\int_0^\tau\sigma^{k_1}\varepsilon^bu_j\left(\frac{\sigma}{\varepsilon^a}\right)d\sigma\right)d\tau$$
$$=\varepsilon^{2b}\int_0^{\bar t}\tau^{k_2}\varepsilon^{ak_2}u_j(\tau)\int_0^\tau\sigma^{k_1}\varepsilon^{k_1a}u_j(\sigma)\varepsilon^{2a}\,d\sigma\,d\tau$$
$$=\varepsilon^{(k_1+k_2+2)a+2b}\int_0^{\bar t}\tau^{k_2}u_j(\tau)\int_0^\tau\sigma^{k_1}u_j(\sigma)\,d\sigma\,d\tau.$$

To prove (ii) let us recall that $S(I, u_\varepsilon, \varepsilon^a t)$ is the solution at time t of the differential equation

$$\dot{S} = S\varepsilon^a \left(X_0 + \sum_{|\nu|=1}^{l} \varepsilon^{b|\nu|} u^\nu X_\nu \right), \qquad S(0) = I.$$

In the same way, $\Gamma(S^0, -\varepsilon^a t)$ is the solution at time $-t$ of

$$\dot{S} = \varepsilon^a S X_0, \qquad S(0) = S^0.$$

Therefore,

$$\Gamma(S(I, u_\varepsilon, \varepsilon^a t), -\varepsilon^a t)$$

can be obtained from $\Gamma(S(I, u, t), -t)$ by substituting $\varepsilon^a X_0$ and $\varepsilon^{a+b|\nu|} X_\nu$, respectively, for X_0 and X_ν. Part (ii) then clearly follows from (11).

5. Proof of Theorems 2 and 3. Clearly, Theorem 2 is the first step of induction of Theorem 3. However, we have enounced it separately in order to point out what can be derived by using only the fact that $\gamma(t)$ is a totally singular arc. Indeed, in Theorem 3 more degenerated situations are considered.

To prove (i) and (ii) of Theorem 2 and (ii) of Theorem 3 we will apply the techniques introduced by Sussmann [21] and summarized in the previous section. To prove (iii) of Theorem 2 and (ii) and (iii) of Theorem 3 we will use the Volterra expansion techniques applied to the approximated system Σ^2. Let us first investigate this approximation of Σ in more detail. Recall that

$$\Sigma^2 \begin{cases} \dot{z}(t) = f_0(z(t)) + \sum_{i=1}^{m} f_i(z(t)) u_i(t) + \frac{1}{2} \sum_{i,j=1}^{m} f_{ij}(z(t)) u_i(t) u_j(t), \\ z(0) = x^0. \end{cases}$$

From [15] we can approximate the solution $z(x^0, u, t)$, or any function of it, $h(z(x^0, u, t))$ (provided that h is a function sufficiently smooth) with a finite Volterra series expansion

$$\begin{aligned} h(z(x^0, u, T)) = w^0(T, x^0) &+ \sum_{j=1}^{m} \int_0^T w_j^1(T, \tau, x^0) u_j(\tau)\, d\tau \\ &+ \frac{1}{2} \sum_{i,j=1}^{m} \int_0^T w_{ij}^{21}(T, \tau, x^0) u_j(\tau) u_i(\tau)\, d\tau \\ &+ \sum_{i,j=1}^{m} \int_0^T \int_0^{\tau_2} w_{ij}^{22}(T, \tau_2, \tau_1, x^0) u_j(\tau_1) u_i(\tau_2)\, d\tau_1\, d\tau_2 \\ &+ O(\|u\|_1^3) + O(\|u\|_2^4). \end{aligned} \tag{12}$$

If the control u is different from zero only in the short interva! of time $[t, t+\varepsilon]$, we can approximate each Volterra kernel by its Taylor polynomial with respect to time and we get

$$w^0(T, x^0) = h[\exp Tf_0(x^0)] = h(\gamma(T)) = w_0(t, \gamma(t)),$$

$$w_j^1(T, \tau, x^0) = \sum_{0 \leqq k \leqq r} \frac{(\tau - t)^k}{k!} \operatorname{ad}_{f_0}^k f_j \cdot w_0(t, \gamma(t)) + O(\varepsilon^{r+1}),$$

$$w_{ij}^{21}(T, \tau, x^0) = \sum_{0 \leqq k \leqq r} \frac{(\tau - t)^k}{k!} \operatorname{ad}_{f_0}^k f_{ij} \cdot w_0(t, \gamma(t)) + O(\varepsilon^{r+1}),$$

$$w_{ij}^{22}(T, \tau_2, \tau_1, x^0) = \sum_{\substack{k_1,k_2\geqq 0\\ k_1+k_2\leqq r}} \frac{(\tau_2-t)^{k_2}}{k_2!}\frac{(\tau_1-t)^{k_1}}{k_1!}\,\mathrm{ad}_{f_0}^{k_1} f_i\cdot \mathrm{ad}_{f_0}^{k_2} f_j\cdot w_0(t, \gamma(t))+O(\varepsilon^{r+1}).$$

As we are considering a singular extremal trajectory, the first and second sums in the right-hand side of (12) vanish. Moreover, by means of the integral identities stated in [4] and using Lemma 4, we obtain the following approximation formula for the solutions of the original system Σ:

$$\begin{aligned} &h(x(x^0, u, T))-h(\gamma(T))\\ &\quad=\frac{1}{2}\sum_{i,j=1}^{m}\sum_{\substack{k_1,k_2\geqq 0\\ k_1+k_2\leqq r}}[\mathrm{ad}_{f_0}^{k_1} f_i, \mathrm{ad}_{f_0}^{k_2} f_j]\cdot w_0(t, \gamma(t))\\ &\qquad\cdot\int_t^{t+\varepsilon}\int_t^{\tau_2}\frac{(\tau_2-t)^{k_2}}{k_2!}\frac{(\tau_1-t)^{k_1}}{k_1!}u_i(\tau_1)u_j(\tau_2)\,d\tau_1\,d\tau_2 \qquad (13)\\ &\qquad+\frac{1}{2}\sum_{i,j=1}^{m}\sum_{\substack{k_1,k_2\geqq 0\\ k_1+k_2\geqq r}}\mathrm{ad}_{f_0}^{k_1} f_i\cdot \mathrm{ad}_{f_0}^{k_2} f_j\cdot w_0(t, \gamma(t))\\ &\qquad\cdot\int_t^{t+\varepsilon}\int_t^{t+\varepsilon}\frac{(\tau_2-t)^{k_2}}{k_2!}\frac{(\tau_1-t)^{k_1}}{k_1!}u_i(\tau_1)u_j(\tau_2)\,d\tau_1\,d\tau_2+\mathscr{R}(u, \varepsilon)\end{aligned}$$

with $\mathscr{R}(u, \varepsilon)=O(\varepsilon^{r+1}\|u\|_1^2+\|u\|_1^3+\|u\|_2^4+\|u\|_3^3)$.

The main idea is now to use controls of the form

$$u_\varepsilon(\tau)=\varepsilon^b u\left(\frac{\tau-t}{\varepsilon}\right)$$

in a short interval of time $[t, t+\varepsilon]$ where u is a suitable fixed control. The resulting output function, which is a function of ε, will be given by its approximation (13) with

$$\mathscr{R}(u_\varepsilon, \varepsilon)=O(\varepsilon^{r+2b+3}+\varepsilon^{3b+3}+\varepsilon^{4b+2}).$$

The values of b and r will then be chosen appropriately so that the leading term of the functional expansion (13) depends on a specific bracket. It is easy to prove the following result.

LEMMA 8.

$$[\mathrm{ad}_{f_0}^{k_1} f_i, \mathrm{ad}_{f_0}^{k_2} f_j]=(-1)^k[\mathrm{ad}_{f_0}^{k_1-k} f_i, \mathrm{ad}_{f_0}^{k_2+k} f_j]+\mathrm{ad}_{f_0} Y$$

for each $k=0, 1, 2, \cdots, k_1$, *with*

$$Y=\sum_{0<h<k}(-1)^h[\mathrm{ad}_{f_0}^{k_1-h} f_i, \mathrm{ad}_{f_0}^{k_2+h-1} f_j].$$

Now let us denote

$$\begin{aligned} H_{ij}^{k_1k_2}(t)&=\langle\lambda(t), [\mathrm{ad}_{f_0}^{k_1} f_i, \mathrm{ad}_{f_0}^{k_2} f_j](t)\rangle\\ &=[\mathrm{ad}_{f_0}^{k_1} f_i, \mathrm{ad}_{f_0}^{k_2} f_j]w_0(t, \gamma(t)).\end{aligned}$$

Proof of part (i) *of Theorem* 3. The proof is by induction on $k_1+k_2=r$. For $r=0$, the statement will follow by Theorem 2. Assume this is true until $r-1\leqq 2s-2$. Using Lemmas 2 and 8 and the induction hypothesis, we get for $r=2k+1\leqq 2s-1$ and hence for $k\leqq s-1$

$$H_{ij}^{k_1k_2}(t)=\pm H_{ij}^{kk+1}(t)=0$$

and

$$H_{ij}^{k_1k_2}(t) = (-1)^{k_1} H_{ij}^{0r}(t) = (-1)^{k_2} H_{ij}^{r0}(t) = (-1)^{k_1+1} H_{ji}^{0r}(t) \quad \text{for } r = 2k \geqq 2.$$

Hence if $i = j$, $H_{ij}^{k_1k_2}(t) = 0$; if $i \neq j$, let us choose $u : [0, 1] \to \mathbf{R}^m$ such that

$$u_k \equiv 0 \quad \text{for } k \neq i, j,$$

$$\int_0^1 u_h(\tau)\, d\tau = \int_0^1 \tau u_h(\tau)\, d\tau = \cdots = \int_0^1 \tau^r u_h(\tau)\, d\tau = 0, \qquad h = i, j \tag{14}$$

and

$$\int_0^1 u_j(\tau) \int_0^\tau (\tau - \sigma)^r u_i(\sigma)\, d\sigma\, d\tau = \operatorname{sign} H_{ji}^{0r}(t).$$

Let us now define $u_\varepsilon = [0, T] \to \mathbf{R}^m$ *by*

$$u_\varepsilon(\tau) = \begin{cases} \varepsilon^{r+2} u\left(\dfrac{\tau - t}{\varepsilon}\right), & \tau \in [t, t + \varepsilon], \\ 0 \quad \text{otherwise.} \end{cases}$$

If ε is sufficiently small, $u_\varepsilon \in \mathcal{U}$. Therefore from (13) and Lemma 7(i) we get

$$\begin{aligned} &h(x^0, u_\varepsilon, T_0) \\ &\quad = h(\gamma(T)) + \frac{1}{2} \sum_{k_1+k_2 \leqq r} \frac{1}{k_1! k_2!} \sum_{l_1, l_2 = 1}^{m} [\mathrm{ad}_{f_0}^{k_1} f_{l_1}, \mathrm{ad}_{f_0}^{k_2} f_{l_2}] w_0(t, \gamma(t)) \varepsilon^{2r+6+k_1+k_2} \\ &\qquad \cdot \int_0^1 u_{l_2}(\tau_2) \tau_2^{k_2}\, d\tau_2 \int_0^{\tau_2} u_{l_1}(\tau_1) \tau_1^{k_1}\, d\tau_1 \\ &\qquad\quad + \frac{1}{2} \sum_{k_1+k_2 \leqq r} \frac{1}{k_1! k_2!} \sum_{l_1, l_2 = 1}^{m} \mathrm{ad}_{f_0}^{k_1} f_{l_1} \cdot \mathrm{ad}_{f_0}^{k_2} f_{l_2} w_0(t, \gamma(t)) \varepsilon^{2r+6+k_1+k_2} \\ &\qquad \cdot \int_0^1 u_{l_1}(\tau_2) \tau_2^{k_2}\, d\tau_2 \int_0^1 u_{l_2}(\tau_1) \tau_1^{k_1}\, d\tau_1 + O(\varepsilon^{3r+7}). \end{aligned}$$

Then the induction hypothesis and the properties (14) of u give

$$\begin{aligned} h(x(x^0, u_\varepsilon, T)) &= h(\gamma(T)) + \frac{\varepsilon^{3r+6}}{2} H_{ij}^{0r}(t) \sum_{k_1+k_2 \leqq r} \frac{1}{k_1! k_2!} \\ &\quad \cdot \left\{ (-1)^{k_1} \int_0^1 \tau^{k_2} u_j(\tau) \int_0^\tau \sigma^{k_1} u_i(\sigma)\, d\sigma\, d\tau \right. \\ &\qquad \left. + (-1)^{k_2+1} \int_0^1 \tau^{k_2} u_i(\tau) \int_0^\tau \sigma^{k_1} u_j(\sigma)\, d\sigma\, d\tau \right\} + O(\varepsilon^{3r+7}). \end{aligned}$$

Integrating by parts gives

$$\begin{aligned} h(x(x^0, u_\varepsilon, T)) &= h(\gamma(T)) + \varepsilon^{3r+6} H_{ij}^{0r}(t) \\ &\quad \cdot \int_0^1 u_j(\tau) \int_0^\tau \sum_{k_1+k_2 = r} (-1)^{k_1} \frac{\tau^{k_2} \sigma^{k_1}}{k_1! k_2!}\, d\sigma\, d\tau + O(\varepsilon^{3r+7}) \end{aligned}$$

or

$$h(x(x^0, u_\varepsilon, T)) = h(\gamma(T)) + \frac{\varepsilon^{3r+6}}{r!} H_{ij}^{0r}(t) \cdot \int_0^1 u_j(\tau) \int_0^\tau (\tau-\sigma)^r u_i(\sigma)\, d\sigma\, d\tau + O(\varepsilon^{3r+7}).$$

Therefore, if $H_{ij}^{0r}(t) \not\equiv 0$ we have obtained that

$$h(x(x^0, u_\varepsilon, T)) - h(\gamma(T)) = M\varepsilon^{3r+6} + O(\varepsilon^{3r+7})$$

with $M < 0$, a contradiction. Hence $H_{ij}^{0r}(t) = 0$.

Proof of part (iii) *of Theorems* 2 *and* 3. Note that Theorem 2(iii) is Theorem 3(iii) with $s = 0$. Let us first prove the following result.

LEMMA 9. *Let* $u : [0, 1] \to \mathbf{R}^m$ *be a control and let* $Iu_i : [0, 1] \to \mathbf{R}$ *be defined by*

$$Iu_i(t) = \int_0^t u_i(\tau)\, d\tau$$

and let

$$Iu_i(1) = \cdots = I^{2s+1}u_i(1) = 0, \qquad i = 1, \cdots, m.$$

Then

$$\sum_{\substack{k_1+k_2=2s+1 \\ k_1,k_2 \geqq 0}} \frac{(-1)^{k_1}}{k_1!k_2!} \int_0^1 \tau^{k_2} u_j(\tau) \int_0^\tau \sigma^{k_1} u_i(\sigma)\, d\sigma\, d\tau = \frac{(-1)^{s+1}}{(2s+1)!} \int_0^1 I^{s+1}u_i(\tau) I^{s+1}u_j(\tau)\, d\tau. \tag{15}$$

Proof. We first note that

$$\sum_{\substack{k_1+k_2=2s+1 \\ k_1,k_2 \geqq 0}} \frac{(-1)^{k_1}}{k_1!k_2!} \tau^{k_2}\sigma^{k_1} = \frac{1}{(2s+1)!}(\tau-\sigma)^{2s+1}.$$

Therefore the left-hand side of (15) can be written as follows:

$$\frac{1}{(2s+1)!} \int_0^1 u_j(\tau) \int_0^\tau (\tau-\sigma)^{2s+1} u_i(\sigma)\, d\sigma\, d\tau$$

or

$$\frac{1}{(2s+1)!} \int_0^1 u_j(\tau) I^{2s+2}u_i(\tau)\, d\tau.$$

On the other hand,

$$\int_0^1 u_j(\tau) I^{2s+2}u_i(\tau)\, d\tau = Iu_j(1) I^{2s+2}u_i(1) - \int_0^1 Iu_j(\tau) I^{2s+1}u_i(\tau)\, d\tau$$

$$= -\int_0^1 Iu_j(\tau) I^{2s+1}u_i(\tau)\, d\tau \cdots$$

$$= (-1)^{s+1} \int_0^1 I^{s+1}u_i(\tau) I^{s+1}u_j(\tau)\, d\tau.$$

The statement is proved.

Let us now turn to the proof of (iii) of Theorem 3. Let u be as in Lemma 9. In particular, this implies that

$$\int_0^1 \tau^k u_i(\tau)\, d\tau = 0, \qquad k = 0, \cdots, 2s+1.$$

Let us define

$$u_\varepsilon(\tau) = \begin{cases} \varepsilon^{2s+3} u\left(\dfrac{\tau - t}{\varepsilon}\right), & \tau \in [t, t+\varepsilon], \\ 0 \quad \text{otherwise.} \end{cases}$$

Using the same arguments as in the proof of (i) of Theorem 3, we obtain

$$h(x(x^0, u_\varepsilon, T_0)) = h(\gamma(T)) + \frac{1}{2} \sum_{\substack{k_1+k_2=2s+1 \\ k_1,k_2 \geqq 0}} \varepsilon^{6s+9} \frac{1}{k_1! k_2!} \sum_{i,j=1}^m H_{ij}^{k_1 k_2}(t) \cdot \int_0^1 \tau^{k_2} u_j(\tau) \int_0^\tau \sigma^{k_1} u_i(\sigma)\, d\sigma\, d\tau + O(\varepsilon^{6s+10}).$$

Therefore

$$h(x(x^0, u_\varepsilon, T)) = h(\gamma(T)) + \frac{\varepsilon^{6s+9}}{2} \sum_{i,j=1}^m H_{ij}^{02s+1}(t) \sum_{\substack{k_1+k_2=2s+1 \\ k_1,k_2 \geqq 0}} \cdot \frac{(-1)^{k_1}}{k_1! k_2!} \int_0^1 \tau^{k_2} u_j(\tau) \int_0^\tau \sigma^{k_1} u_i(\sigma)\, d\sigma\, d\tau + O(\varepsilon^{6s+10}).$$

From Lemma 9

$$h(x(x^0, u_\varepsilon, T)) - h(\gamma(T)) = \frac{(-1)^{s+1}}{(2s+1)!} \varepsilon^{6s+9} \sum_{i,j=1}^m H_{ij}^{02s+1}(t) \cdot \int_0^1 I^{s+1} u_i(\tau) I^{s+1} u_j(\tau)\, d\tau + O(\varepsilon^{6s+10})$$

or

$$h(x(x^0, u_\varepsilon, T)) - h(\gamma(T)) = \frac{\varepsilon^{6s+9}}{(2s+1)!} \sum_{i,j=1}^m H_{ij}^{s+1\, s}(t) \cdot \int_0^1 I^{s+1} u_i(\tau) I^{s+1} u_j(\tau)\, d\tau + O(\varepsilon^{6s+10}).$$

If ε is sufficiently small $u_\varepsilon \in \mathcal{U}$, therefore with the chosen control u we must have

$$\sum_{i,j=1}^m H_{ij}^{s+1\, s}(t) \int_0^1 I^{s+1} u_i(\tau) I^{s+1} u_j(\tau)\, d\tau \geqq 0. \tag{16}$$

Lemma 8 and part (i) imply

$$H_{ij}^{s+1\, s}(t) = -H_{ij}^{s\, s+1}(t) = H_{ij}^{s+1\, s}(t).$$

Let us now prove that the matrix

$$Q^s(t) = ((H_{ij}^{s+1\,s}(t)))_{1\leqq j\leqq m, 1\leqq j\leqq m}$$

is a nonnegative matrix.

If the matrix $Q^s(t)$ is not positive semidefinite the set

$$\left\{\mu\in\Omega\colon \sum_{i,j=1}^{m} H_{ij}^{s+1\,s}(t)\mu_i\mu_j<0\right\}$$

contains an open cone. Therefore there exists a C^∞ map $p\colon[0,1]\to\mathbf{R}^m$ such that

$$\sum_{i,j=1}^{m} H_{ij}^{s+1\,s}(t)p_i(\tau)p_j(\tau)<0 \quad \forall\tau\in(0,1) \tag{17}$$

and $p_i(0)=p_i^{(1)}(0)=\cdots=p_i^{(s+1)}(0)=p_i(1)=p_i^{(1)}(1)=\cdots=p_i^{(s+1)}(1)=0$ for $i=1,\cdots,m$.

Let us choose $u=p^{(s+1)}$. This control u satisfies the hypothesis of Lemma 9 and

$$\sum_{i,j=1}^{m} H_{ij}^{s+1\,s}(t)\int_0^1 I^{s+1}u_i(\tau)I^{s+1}u_j(\tau)\,d\tau=\sum_{i,j=1}^{m} H_{ij}^{s+1\,s}(t)\int_0^1 p_i(\tau)p_j(\tau)\,d\tau,$$

which is negative from (17) and then contradicts (16).

Proof of parts (i) *and* (ii) *of Theorem* 2 *and part* (ii) *of Theorem* 3. From Lemma 2 it is sufficient to prove the statements for $k=0$. Let us consider the approximating system Σ_r^l of Σ with $l=3$ and $r=4s+4$, $s\geqq 0$. Let $Y_i\in X$ and let $c=(0,\cdots,c_i,0,\cdots,0)$ be given by Theorem 4. We have

$$\Gamma(S(I,\xi(c),\bar t),-\bar t)=\exp\left[\sum_{j=1,\cdots,N_r} q(j,\bar u,\bar t)Y_j+c_iY_i\right].$$

Let us define $\xi_\varepsilon(c)$ by

$$\xi_\varepsilon(c)(\tau)=\begin{cases}\varepsilon^{s+1}\xi(c)\left(\dfrac{\tau}{\varepsilon}\right) & \text{for } t\in[0,\varepsilon\bar t],\\ 0 \quad \text{otherwise}\end{cases}$$

from the definition of ξ_ε, $\|\xi_\varepsilon(c)\|_4^4=O(\varepsilon^{4s+5})$. Therefore from Lemma 6 it follows that

$$\|\mathrm{ev}_{\gamma(t)}\,\Gamma(S(I,\xi_\varepsilon(c),\varepsilon\bar t),-\varepsilon\bar t)x-\exp(-\varepsilon\bar t f_0)x(\gamma(t),\xi_\varepsilon(c),\varepsilon\bar t)\|\leqq H\varepsilon^{4s+5}.$$

As we are considering the approximation Σ_r^3 of Σ, if $Y_j\in X$ is even and $\omega(Y_j)>2$, then $\omega(Y_j)\geqq 4$ and $\|Y_j\|\geqq 2$. Now from Lemma 7 we obtain

$$\Gamma(S(I,\xi_\varepsilon(c),\varepsilon\bar t),-\varepsilon\bar t)=\exp\left(\sum_{\substack{\beta_j=2\\ Y_j\text{ even}\\ \alpha_j\leqq 2s+2}} q(j,\bar u,\bar t)Y_j\varepsilon^{2(s+1)+\alpha_j}+c_iY_i\varepsilon^{\beta_i(s+1)+\alpha_i}+O(\varepsilon^{4s+5})\right).$$

Let us investigate for which j, $q(j,\bar u,\bar t)$ may be different from zero. If $\beta_j=2$ and $\alpha_j=1$, then $Y_j=X_\nu$ for some $\nu=(0,\cdots,2,0,\cdots,0)$. In the case where $s=0$, $\beta_j=2$, and $1<\alpha_j\leqq 2s+2$, we give $Y_j=[X_{l_1},X_{l_2}]$ for some $l_1,l_2\in\{1,\cdots,m\}$. Hence Y_j cannot be **even** if $\alpha_j>1$. Now if $s\geqq 1$ the Y_j's even with $\beta_j=2$ and $\alpha_j\leqq 2s+2$ can be either of type $\mathrm{ad}_{X_0}^k X_\nu$ for some $\nu=(0,\cdots,2,0,\cdots,0)$, $0\leqq k\leqq 2s+1$ or of type $[\mathrm{ad}_{X_0}^{k_1}X_{l_1},\mathrm{ad}_{X_0}^{k_2}X_{l_2}]$ with $k_1+k_2\leqq 2s$. Therefore using part (i) of Theorem 3 (if $s\geqq 1$) and the fact that γ is a singular trajectory, we obtain

$$\mathrm{ev}\ Y_j w_0(t,\gamma(t))=0$$

if $\omega(Y_j)=\beta_j=2$, $\|Y_j\|=\alpha_j\leqq 2s+2$, and Y_j even.

Now let

$$\eta_\varepsilon(c)(\tau)=\begin{cases}\xi_\varepsilon(c)(\tau-t) & \text{for } t\in[t,\, t+\varepsilon\bar{t}],\\ 0 & \text{otherwise.}\end{cases}$$

We get

$$\begin{aligned}h(x(x^0,\,\eta_\varepsilon(c),\,T)) &= w_0(t,\,\exp(-\varepsilon\bar{t}f_0)\cdot x(\gamma(t),\,\xi_\varepsilon(c),\,\varepsilon\bar{t}))\\ &= w_0(t,\,ev_{\gamma(t)}\Gamma(S(I,\,\xi_\varepsilon(c),\,\varepsilon\bar{t}),\,-\varepsilon\bar{t})x+O(\varepsilon^{4s+5})).\end{aligned}$$

Finally, the equality

$$\sum_{\substack{Y_j\ \text{even}\\ \alpha_j\leqq 2s+2}} q(j,\,\bar{u},\,\bar{t})\,Y_j\varepsilon^{2(s+1)+\alpha_j}=O(\varepsilon^{2s+3})$$

allows us to write

$$h(x(x^0,\,\eta_\varepsilon(c),\,T))=w_0(t,\,\gamma(t))+\varepsilon^{\beta_i(s+1)+\alpha_i}c_i\ \text{ev}\ Y_i\cdot w_0(t,\,\gamma(t))+O(\varepsilon^{4s+5}).$$

Therefore, if

$$\phi=\text{ev}\ Y_i \quad\text{and}\quad \beta_i(s+1)+\alpha_i<4s+5 \tag{18}$$

it follows that

$$\phi\cdot w_0(t,\,\gamma(t))=0 \quad \text{for } t\in[0,\,T].$$

We have proven together parts (i) and (ii) of Theorem 2 and part (ii) of Theorem 3. Indeed, assuming $s=0$, from (18) we obtain

$$\beta_i+\alpha_i<5,$$

which gives two alternatives: either

$\beta_i=\alpha_i=2$, which proves part (i) of Theorem 2, or

$\beta_i=3$ and $\alpha_i=1$, which proves part (ii) of Theorem 2.

Assuming now that $s\geqq 1$, then

$\beta_i=3$ and $\alpha_i\leqq s+1$, which proves part (ii) of Theorem 3.

6. Final comments. In this paper for the sake of simplicity, we have considered only totally singular arcs relative to C^∞ controls.

The same techniques can be used in an obvious way if the control is piecewise C^∞ and only a subarc of the extremal arc is singular. If the system is affine with respect to the control, i.e.,

$$\dot{x}=f_0(x)+\sum_{i=1}^{m}u_if_i(x)$$

all the trajectories are totally singular. In this case (ii) of Theorem 3 can be improved as follows.

THEOREM 5. *Let $\gamma(t)$ be a minimal trajectory. If*

$$[\text{ad}_{f_0}^{k+1}f_i,\,\text{ad}_{f_0}^{k}f_j]\cdot w_0(t,\,\gamma(t))=0 \text{ for } t\in[0,\,T],$$

$$i,\,j=1,\,\cdots,\,m \text{ and } k=0,\,\cdots,\,s-1,\ s\geqq 1,$$

then

$$\text{ad}_{f_0}^{k}\phi w_0(t,\,\gamma(t))=0 \quad \text{for } t\in[0,\,T],\ k\geqq 0 \text{ and } \phi=ev\Lambda\ \omega(\Lambda)=3 \text{ and } \|\Lambda\|\leqq s+3.$$

Proof. The proof is quite similar to that of (ii) of Theorem 3. In this case $\Sigma = \Sigma^l$ for all $l \geqq 1$ so that (10) becomes $\|\mathrm{ev}_y\, \Gamma(S(I, u, t), -t) \cdot x - \exp(-tf_0) \cdot x(y, u, t)\| \leqq Ht^{r+1}$. Let us now consider Σ_r with $r = 4s+1$ and let Y_i and $c = (0, \cdots, c_i, 0, \cdots, 0)$ be such that

$$\omega(Y_i) = 3, \quad \|Y_i\| \leqq s+3, \quad \mathrm{ev}\ Y_i = \Phi$$

and

$$\Gamma(S(I, \xi(c), \bar{t}), -\bar{t}) = \exp\left[\sum_{j=1,\cdots,N_r} q(j, \bar{u}, \bar{t})\, Y_j + c_i Y_i\right].$$

As previously, let us now define the function $\xi_\varepsilon(c)$ by

$$\xi_\varepsilon(c)(\tau) = \begin{cases} \varepsilon^{s-1/2}\xi(c)\left(\dfrac{\tau}{\varepsilon}\right) & \text{for } \tau \in [0, \varepsilon], \\ 0 \quad \text{otherwise.} \end{cases}$$

If $\beta_j = 2$ and $\alpha_j \geqq 2s+3$, then

$$q(j, \bar{u}_\varepsilon, \varepsilon\bar{t}) = O(\varepsilon^{4s+2}).$$

If Y_j is **even** with $\beta_j > 2$, then $\beta_j \geqq 4$, $\alpha_j = \|Y_j\| \geqq 4$, and

$$q(j, \bar{u}_\varepsilon, \varepsilon\bar{t}) = O(\varepsilon^{4s+2}).$$

Therefore

$$\Gamma(S(I, \xi_\varepsilon(b), \varepsilon\bar{t}) - \varepsilon\bar{t}) = \exp\left[\sum_{\substack{\beta_j = 2 \\ Y_j \text{ even} \\ \alpha_j \leqq 2s+2}} q(j, \bar{u}, \bar{t})\varepsilon^{2s-1+\alpha_j} Y_j + c_i\varepsilon^{3s-3/2+\alpha_i} Y_i + O(\varepsilon^{4s+2})\right].$$

On the other hand,

$$3s - \tfrac{3}{2} + \alpha_i \leqq 4s + \tfrac{3}{2}$$

and we apply the same arguments to prove (ii) of Theorem 3.

Acknowledgment. The kind hospitality of Professor Peter Crouch is gratefully acknowledged.

REFERENCES

[1] A. A. AGRACEV, *A second order necessary condition for optimality in the general nonlinear case*, Math. USSR-Sb., 31 (1977), pp. 493-506.

[2] D. J. BELL AND D. H. JACOBSON, *Singular Optimal Control Problems*, Academic Press, London, 1975.

[3] R. W. BROCKETT, *Lie theory, functional expansions and necessary conditions in optimal control*, in Mathematical Control Theory, W. A. Coppel, ed., Lecture Notes in Mathematics, 680, Springer-Verlag, Berlin, 1978, pp. 68-76.

[4] P. E. CROUCH AND F. LAMNABHI-LAGARRIGUE, *Algebraic and multiple integral identities*, Acta Mathematica Applicandae, to appear.

[5] P. E. CROUCH, *Graded and nilpotent approximations of input-output systems*, in Linear and Nonlinear Mathematical Control Theory, Rendiconti Del Seminario Matematico, Torino, Italy, 1987, pp. 1-54.

[6] V. GABASOV AND F. M. KIRILLOVA, *High order necessary conditions for optimality*, SIAM J. Control. 10 (1972), pp. 127-168.

[7] B. S. GOH, *The second variation for the singular Bolza problem*, SIAM J. Control Optim., 4 (1966), pp. 309-325.

[8] V. GOROKHOVIK, *High order necessary optimality conditions for control problems with terminal constraints*, Optimal Control Appl. Methods, 4 (1983), pp. 103-127.

[9] D. H. JACOBSON AND J. L. SPEYER, *Necessary and sufficient conditions for optimality for singular control problems: a limit approach*, J. Math. Anal. Appl., 34 (1971), pp. 239–266.

[10] M. A. KAZEMI-DEHKORDI, *Necessary conditions for optimality of singular controls*, J. Optim. Theory Appl., 43 (1984), pp. 630–637.

[11] H. J. KELLEY, R. E. KOPP, AND H. G. MOYER, *Singular extremals*, in Topics in Optimization, G. Leitman, ed., Academic Press, New York, 1967.

[12] H. W. KNOBLOCH, *Higher Order Necessary Conditions in Optimal Control Theory*, Lecture Notes in Control and Information Science, 34, Springer-Verlag, Berlin, 1985.

[13] A. J. KRENER, *The higher order maximum principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256–293.

[14] F. LAMNABHI-LAGARRIGUE, *Series de Volterra et commande optimale singulière*, Thèse d'Etat, Université Paris XI, Paris, France, 1985.

[15] ———, *Singular optimal control problems: on the order of a singular arc*, Systems Control Lett., 9 (1987), pp. 173–182.

[16] F. LAMNABHI-LAGARRIGUE AND M. ROSSET, *Nonlinear heat transfer control problems*, preprint, 1988.

[17] L. PONTRYAGIN, V. BOLTYANSKII, R. GAMKRELIDZE, AND E. MICHTCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[18] G. STEFANI, *A sufficient condition for extremality*, in Proc. INRIA Conference, Antibes, Lecture Notes in Control and Information Science, 3, Springer-Verlag, Berlin, 1988, pp. 270–281.

[19] ———, *On Volterra series approximations*, in Proc. Nonlinear Control Conference, Nantes, 1988, Lecture Notes in Control and Information Science, Springer-Verlag, Berlin, to appear.

[20] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

[21] H. J. SUSSMANN, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

[22] N. X. VINH, *Optimal Trajections in Atmospheric Flight*, Studies in Aeronautics 2, Elsevier, Amsterdam, 1981.

[23] K. WAGNER, *Uber den Steuerbarkeitsbegriff bei nichtlinearen Kontrollsystemen*, Arch. Math. (Basel), 47 (1986), pp. 29–40.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 841-850, July 1990

005

STOPPING RULES FOR A RANDOM OPTIMIZATION METHOD*

C. C. Y. DOREA†

Abstract. A stochastic algorithm for estimating the global minimum of a function is described and two types of stopping rules are derived. The first is based on the estimation of the region of attraction of the global minimum; the second is based on the existence of the asymptotic distribution of properly normalized estimators.

Key words. random optimization, sequential methods, stopping rules

AMS(MOS) subject classifications. 65K10, 60F99, 62L15

1. Introduction. Global optimization problems have attracted a great deal of attention from researchers in recent years. Aside from deterministic approaches, stochastic methods have been developed. It is generally assumed that a prescribed domain $\Omega \subset R^d$, containing the global minimum point, is given in advance and the problem is to find

$$\mathbf{y}_0 = \min_{x \in \Omega} \{f(x)\} \qquad (\text{or } \max_{x \in \Omega} \{f(x)\}) \tag{1}$$

where $f: \Omega \to R$ satisfies some regularity conditions and the global minimum $\mathbf{y}_0$ is assumed to be finite. The simplest stochastic method, the pure random search method, starts from a random sample of points, $(\xi_1, \cdots \xi_n)$, drawn from a uniform distribution in the domain Ω and yields a candidate solution $Y_n = \min\{f(\xi_1), \cdots, f(\xi_n)\}$ with some asymptotic probabilistic qualities. In fact, for $B \subset \Omega$, the probability that a uniform sample of size n contains at least one point of B is equal to

$$P(B) = 1 - \left[1 - \frac{\mathbf{m}(B)}{\mathbf{m}(\Omega)}\right]^n \tag{2}$$

where $\mathbf{m}$ is the Lebesgue measure of R^d. Thus, if $\mathbf{x}_0$ is such that $f(\mathbf{x}_0) = \mathbf{y}_0 < f(x)$ for $x \neq \mathbf{x}_0$ and the measure of the set

$$B(\varepsilon, \mathbf{x}_0) = \{x: x \in \Omega, \|x - \mathbf{x}_0\| < \varepsilon\} \tag{3}$$

is positive, then a point within distance ε from the global minimum point $\mathbf{x}_0$ will be found with a probability approaching 1 as n increases. Also, as shown by Solis-Wets [9], under a very general setting, only mild conditions need to be imposed on f to obtain almost sure convergence,

$$P(\lim_{n \to \infty} Y_n = \mathbf{y}_0) = 1 \qquad (Y_n \xrightarrow[n]{\text{a.s.}} \mathbf{y}_0).$$

Furthermore, an application of statistics of extreme values leads us to the existence of limiting theorems of the type,

$$\lim_{n \to \infty} P(Y_n \leqq \mathbf{y}_0 + a_n y) = 1 - \exp(-y^\alpha) \tag{4}$$

where $\alpha > 0$ is the shape parameter of the limiting distribution and $a_n \downarrow 0$ are the norming constants (see de Haan [6] or Dorea [4]).

* Received by the editors May 9, 1988; accepted for publication (in revised form) September 30, 1989.
† Instituto de Ciencias Exatas, Universidade de Brasília, 70919 Brasilia-DF, Brasil and Departmento de Estastica IMECC, Universidade Estadual de Campinas, Caixa Postale 6065, 13081 Campinas, S.P., Brasil.
This work was partially supported by Conselho Nacional de Pesquisa, Brasil.

The more refined methods, such as clustering procedures, multistart techniques, or Bayesian methods start from a random sample of points drawn from Ω. This sampling phase is then followed by a local phase in which the sample is manipulated to yield a possible solution of the problem. The algorithm will then return to the sampling phase with information of the best points obtained, and will continue, possibly with a recommendation for a modified sampling distribution. The procedure will continue until some prescribed criteria of optimality are satisfied. In Devroye [3], for f satisfying Lipschitz type conditions and for a progressive random search procedure, convergence efficiency of the algorithm is studied through the ratio

$$u_n = \frac{P(Y_1 > \mathbf{y}_0 + \varepsilon)}{P(Y_n > \mathbf{y}_0 + \varepsilon)}.$$

It is shown that $u_n \to \infty$ as $n \to \infty$. In the clustering method studied by Boender et al. [1] the termination criteria are based on the existence of new local minima points satisfying certain measures associated with the clustering procedure. For the multistart technique, Rinnooy-Kan and Timmer [7] propose a stopping rule, based on a Bayesian estimate, of the number of local minima points and the relative size of each region of attraction of the local minimum points. They all offer an asymptotic guarantee of reaching the global minimum $\mathbf{y}_0$.

The question of developing stopping rules that provide some explicit probabilistic information on the quality of the proposed solution still remains to be satisfactorily solved, even for the pure random search method.

In this note we propose stopping rules for the sequential methods described below.

ALGORITHM A. Let $\xi_1, \xi_2, \cdots$ be independent and identically distributed random vectors with a common distribution G on Ω. Let $(X_1, Y_1), (X_2, Y_2) \cdots$ be defined by

Step 1. $X_1 = \xi_1$ and $Y_1 = f(\xi_1)$.

Step $k+1$. Having defined (X_k, Y_k), let (X_{k+1}, Y_{k+1}) be defined by,

(i) $X_{k+1} = \xi_{k+1}$ and $Y_{k+1} = f(\xi_{k+1})$ if $f(\xi_{k+1}) \leqq Y_k$.

(ii) $X_{k+1} = X_k$ and $Y_{k+1} = Y_k$ otherwise.

Note that the uniform sampling on Ω has been replaced by sampling under a distribution G. Our next algorithm suits the cases in which Ω is not a "nice set" and it is convenient to sample in a larger set $\Omega_0 \supset \Omega$:

ALGORITHM B. Let $\xi_1, \xi_2, \cdots$ be independent and identically distributed random vectors with a common distribution G on Ω_0. Let $(X_0, Y_0), (X_1, Y_1), \cdots$ be defined by

Step 0. $X_0 = U$ and $Y_0 = f(U)$, where U is uniformly distributed over Ω.

Step $k+1$. Having defined (X_k, Y_k), let (X_{k+1}, Y_{k+1}) be defined by

(i) $X_{k+1} = \xi_{k+1}$ and $Y_{k+1} = f(\xi_{k+1})$ if $\xi_{k+1} \in \Omega$ and $f(\xi_{k+1}) < Y_k$.

(ii) $X_{k+1} = X_k$ and $Y_{k+1} = Y_k$ otherwise.

Two types of stopping rules will be studied. The first (Rules 1a and 1b) concerns the ε-region of attraction of the global minimum $\mathbf{y}_0$:

$$A(\varepsilon) = \{x : x \in \Omega, f(x) \leqq \mathbf{y}_0 + \varepsilon\} \tag{5}$$

and the second (Rule 2) concerns the ε-region of attraction of the global minimum point $\mathbf{x}_0$ (see (3)).

Stopping Rule 1. For given $\varepsilon > 0$ and $0 < \beta < 1$ terminate Algorithm A

1a) for n such that

$$n \geqq \log \beta / \log\left(1 - \frac{\rho_n(\varepsilon)}{n}\right) \tag{6}$$

1b) whenever a value of Y_n is repeated for the next m steps, that is, $Y_n = Y_{n+j}, j = 0, 1, \cdots, m$, and m satisfies

$$m \geqq \frac{\log \beta}{\log\left(1 - \frac{\rho_n(\varepsilon)}{n}\right)} - n, \tag{7}$$

where

$$\rho_n(\varepsilon) = \sup\{k : \tau_k > 0,\ Y_{\tau_k} \leqq Y_n + \varepsilon\}, \tag{8}$$

and for $j = 1, 2, \cdots, n-1$ we define

$$\begin{aligned} \tau_{j+1} = \tau_{j+1}(n) &= \sup\{k : 1 \leqq k \leqq \tau_j,\ Y_k \neq Y_{\tau_j}\} \\ &= 0 \text{ if } \{k : 1 \leqq k \leqq \tau_j,\ Y_k \neq Y_{\tau_j}\} = \varnothing \end{aligned} \tag{9}$$

with $\tau_1(n) = n$.

In § 2 we will show that, if Rule 1a is applied, then

$$P(A(\varepsilon)) = P(|Y_n - \mathbf{y}_0| \leqq \varepsilon) \geqq 1 - \beta. \tag{10}$$

That is, the ε-region of attraction $A(\varepsilon)$ of $\mathbf{y}_0$ has been attained with a prescribed probability $1-\beta$. Alternatively, if Rule 1b is applied we have,

$$P(|Y_n - \mathbf{y}_0| \leqq \varepsilon \,|\, Y_n = Y_{n+j}, j = 0, 1, \cdots, m) \geqq 1 - \beta. \tag{11}$$

Note that at step l we have $\tau_1 = l$ and $Y_l = Y_{l-1} = \cdots = Y_{\tau_2 - 1} < Y_{\tau_2} = Y_{\tau_2+1} = \cdots = Y_{\tau_3 - 1} < Y_{\tau_3} \cdots$, so that Rule 1b could be replaced by

$$\tau_1 - \tau_2 + 1 \geqq \log \beta / \log\left(1 - \frac{\rho_{\tau_2 - 1}}{\tau_2 - 1}\right)$$

Our next stopping rule is based on the asymptotic distribution $1 - \exp(-y^\alpha)$ of the Y_n's properly normalized and under the assumption of the following condition.

CONDITION 1. *Assume that $\Omega \subset R$ with $\mathbf{m}(\Omega) > 0$ and that G is the uniform distribution on Ω. Moreover,*

(a) *There exists a unique interior point $\mathbf{x}_0$ of Ω such that $f(\mathbf{x}_0) = \mathbf{y}_0$.*

(b) *There exists a positive function $\nu(t)$, $t > 0$, and a constant $\alpha > 0$ such that for all $x > 0$, $\lim_{t \downarrow 0} (\nu(tx)/\nu(t)) = x^{1/\alpha}$ and the following limit*

$$R(z) = \lim_{t \downarrow 0} \frac{f(\mathbf{x}_0 + tz) - \mathbf{y}_0}{\nu(t)}$$

exists and is strictly positive and finite for all $z \neq 0$.

Although Condition 1(b) is not easy to verify, the following example shows that it is not too restrictive. In fact, f need not be differentiable. Let $\Omega = [-1, 1]^2$ and $f(x, y) = \max(|x|, |y|)$. Then Condition 1(b) is satisfied by taking $\nu(t) = t$.

Stopping Rule 2. For given $\varepsilon > 0$ and $0 < \beta < 1$, terminate Algorithm A for n such that

$$n \geqq -\frac{\mathbf{m}(\Omega) \log \beta}{\varepsilon} \tag{12}$$

It will follow from Theorem 1 (see § 2) that if (12) is satisfied then,

$$P(|X_n - \mathbf{x}_0| \leqq \varepsilon) \approx 1 - \beta. \tag{13}$$

In addition, under certain assumptions, we can also conclude that

$$P(|Y_n - \mathbf{y_0}| \leqq \varepsilon^{1/\alpha}) \geqslant 1-\beta. \tag{14}$$

The discussion for the situation $\Omega \subset R^d$ will be carried out in § 2. In this case, (12) and (13) are, respectively, replaced by,

$$n \geqq -\frac{\mathbf{m}(\Omega)\log\beta}{\varepsilon^d} \tag{15}$$

$$\text{and} \quad P((X_n - \mathbf{x_0}) \in [-\varepsilon, \varepsilon]^d) \approx 1-\beta. \tag{16}$$

Finally, in § 3, we present some numerical and analytical comparisons between the rules and a discussion of how these rules can be adapted for Algorithm B.

2. Stopping rules for Algorithm A. The stopping rules 1a and 1b are derived by first observing that under Algorithm A we have $Y_n = \min\{Y_1, \cdots, Y_n\}$. And if $Z_j = f(\xi_j)$ for $j = 1, \cdots, n$ we also have $Y_n = \min\{Z_1, \cdots, Z_n\}$ where the Z_j's are independent and identically distributed random variables with a common distribution given by

$$F(x) = P(Z_1 \leqq x) = P(f(\xi_1) \leqq x) = \int_{\{u: f(u) \leqq x\}} dG(u). \tag{17}$$

Moreover, for a given $\varepsilon > 0$

$$\begin{aligned} P(Y_n \leqq \mathbf{y_0} + \varepsilon) &= 1 - P(Y_n > \mathbf{y_0} + \varepsilon) = 1 - (P(Z_1 > \mathbf{y_0} + \varepsilon))^n \\ &= 1 - (1 - F(\mathbf{y_0} + \varepsilon))^n = 1 - (1 - p_\varepsilon)^n. \end{aligned}$$

Hence, if $0 < \beta < 1$ and p_ε is known, we have

$$P(|Y_n - \mathbf{y_0}| \leqq \varepsilon) \geqq 1 - \beta \tag{18}$$

provided that $n \geqq \log\beta / \log(1 - p_\varepsilon)$. Similarly, if a certain value Y_n is repeated in the next m steps of the algorithm, we have

$$P(Y_n = Y_{n+j},\ j = 1, \cdots, m) = P(Z_{n+j} > Y_n,\ j = 1, \cdots, m).$$

Note that Y_n has a distribution given by

$$H(x) = P(Y_n \leqq x) = 1 - (1 - F(x))^n,$$

F being the common distribution of the independent and identically distributed random variables $Z_{n+1}, \cdots, Z_{n+m}$. Since the random vector $(Z_{n+1}, \cdots, Z_{n+m})$ is independent of Y_n, we have

$$\begin{aligned} P(Y_{n+j} = Y_n, j = 1, \cdots, m) &= \int (1 - F(x))^m \, dH(x) \\ &= n \int (1 - F(x))^{n+m-1} \, dF(x) = \frac{n}{m+n}. \end{aligned}$$

Analogous arguments show that

$$\begin{aligned} &P(Y_{n+j} = Y_{n,j} = 1, \cdots, m,\ Y_n > \mathbf{y_0} + \varepsilon) \\ &= n \int_{\mathbf{y_0}+\varepsilon}^{\infty} (1 - F(u))^{m+n-1} \, dF(u) = \frac{n}{m+n}(1 - p_\varepsilon)^{m+n}. \end{aligned}$$

Hence, if $m \geqq (\log\beta / \log(1 - p_\varepsilon)) - n$ we have

$$P(|Y_n - \mathbf{y_0}| \leqq \varepsilon \,|\, Y_n = Y_{n+j}, j = 0, 1, \cdots, m) = 1 - (1 - p_\varepsilon)^{n+m} \geqq 1 - \beta. \tag{19}$$

Since p_ε is unknown, our stopping rule makes use of the quantity $\rho_n(\varepsilon)/n$ defined by (8). The reason for such replacement can be understood from the following derivation of an estimator of p_ε. Let $(Z_{(1,n)}, Z_{(2,n)}, \cdots, Z_{(n,n)})$ denote the ordered sample (order statistics) of $(Z_1, \cdots, Z_n)$. That is, $Z_{(1,n)} = \min\{Z_1, \cdots, Z_n\}$; $Z_{(2,n)}$ denote the second lowest value of the sample; $\cdots$; and $Z_{(n,n)} = \max\{Z_1, \cdots, Z_n\}$. For $\varepsilon > 0$ define

$$\gamma_n(\varepsilon) = \sup\{k : Z_{(k,n)} \leqq \mathbf{y}_0 + \varepsilon\}. \tag{20}$$

Since $Z_1, \cdots, Z_n$ are independent and identically distributed random variables and $P(Z_1 \leqq \mathbf{y}_0 + \varepsilon) = p_\varepsilon$ we have $\gamma_n(\varepsilon)$ binomially distributed with parameters n and p_ε. If $p_\varepsilon > 0$, it follows from the strong law of large numbers that $\gamma_n(\varepsilon)/n \xrightarrow[n]{\text{a.s.}} p_\varepsilon$. Since $\mathbf{y}_0$ is unknown and the $Z_{(k,n)}$'s are not recorded by our algorithm, we approximate $\mathbf{y}_0$ by Y_n and $Z_{(k,n)}$ by Y_{τ_k}. That is, we approximate $\gamma_n(\varepsilon)$ by $\rho_n(\varepsilon)$. The following proposition justifies the proposed rules.

PROPOSITION 1. *Let $\varepsilon > 0$ and $0 < \beta < 1$. Assume that for all $\varepsilon > 0$ we have $p_\varepsilon = P(f(x) \leqq \mathbf{y}_0 + \varepsilon) > 0$ and $P(f(x) \leqq \mathbf{y}_0) = 0$. Then for ε small we have*

$$\lim_{\eta \downarrow 0} \lim_{n \to \infty} P\left(\frac{\rho_n(\varepsilon)}{n} \leqq p_{\varepsilon+\eta}\right) = 1, \tag{21}$$

$$\lim_{\eta \downarrow 0} \lim_{n \to \infty} P(\delta_n(\varepsilon) \geqq \hat{\delta}(\varepsilon + \eta)) = 1 \tag{22}$$

and

$$\lim_{\eta \downarrow 0} \lim_{n \to \infty} P(|Y_n - \mathbf{y}_0| \leqq \varepsilon, \delta_n(\varepsilon) \geqq \hat{\delta}(\varepsilon + \eta)) = 1, \tag{23}$$

where

$$\delta_n(\varepsilon) = \frac{\log \beta}{\log\left(1 - \dfrac{\rho_n(\varepsilon)}{n}\right)} \quad \text{and} \quad \hat{\delta}(\gamma) = \frac{\log \beta}{\log(1 - p_\gamma)}.$$

Proof. First note that $p_\varepsilon = F(\mathbf{y}_0 + \varepsilon)$ with F defined by (17). The right continuity of the distribution F gives $\lim_{\eta \downarrow 0} p_{\varepsilon+\eta} = p_\varepsilon$. Moreover, since $p_\varepsilon > 0$ for $\varepsilon > 0$ and $F(\mathbf{y}_0) = 0$, we have p_ε strictly increasing on ε for ε sufficiently small.

Now let $\varepsilon > 0$ and $\eta > 0$. We may assume that $p_{\varepsilon+\eta} > p_{\varepsilon+\eta_1}$ for $0 < \eta_1 < \eta$. Let $\eta_2 > 0$ such that $p_{\varepsilon+\eta} \geqq p_{\varepsilon+\eta_1} + \eta_2$. It follows that

$$\begin{aligned} P\left(\frac{\rho_n(\varepsilon)}{n} \leqq p_{\varepsilon+\eta}\right) &\geqq P\left(\frac{\rho_n(\varepsilon)}{n} \leqq p_{\varepsilon+\eta_1} + \eta_2\right) \\ &\geqq P\left(\frac{\rho_n(\varepsilon)}{n} \leqq \frac{\gamma_n(\varepsilon+\eta_1)}{n}, \left|\frac{\gamma_n(\varepsilon+\eta_1)}{n} - p_{\varepsilon+\eta_1}\right| < \eta_2\right). \end{aligned}$$

From the definitions of $\rho_n(\varepsilon)$ and $\gamma_n(\varepsilon)$, we have $(Y_n \leqq \mathbf{y}_0 + \eta_1) \subset (\rho_n(\varepsilon) \leqq \gamma_n(\varepsilon + \eta_1))$ and we can write

$$P\left(\frac{\rho_n(\varepsilon)}{n} \leqq p_{\varepsilon+\eta}\right) \geqq P\left(\left|\frac{\gamma_n(\varepsilon+\eta_1)}{n} - p_{\varepsilon+\eta_1}\right| < \eta_2, Y_n \leqq \mathbf{y}_0 + \eta_1\right).$$

Since $(\gamma_n(\varepsilon + \eta_1)/n) \xrightarrow[n]{\text{a.s.}} p_{\varepsilon+\eta_1}$ and $Y_n \xrightarrow[n]{\text{a.s.}} \mathbf{y}_0$, we have (21) by taking the proper limits.

From the definitions of $\delta_n(\varepsilon)$ and $\hat{\delta}(\varepsilon+\eta)$, we have (22) and (23) by using (21) and the fact that $Y_n \xrightarrow[n]{\text{a.s.}} \mathbf{y_0}$. □

For stopping rule 2, it is assumed that Condition 1 holds. In this case we have from Dorea [5] that there exist norming constants $a_n \downarrow 0$ such that, for $y>0$

$$\lim_{n\to\infty} P(Y_n \leqq \mathbf{y_0} + a_n y) = 1 - \exp(-y^{\alpha}). \tag{24}$$

For $R(z)$ defined by Condition 1 we have $R(z) = z^{1/\alpha}R(1)$ for $z>0$ and $R(z) = |z|^{1/\alpha}R(-1)$ for $z<0$. Moreover, we may choose $\nu(t)$ so that for

$$k_1 = \left(\frac{1}{R(-1)}\right)^{\alpha} \quad \text{and} \quad k_2 = \left(\frac{1}{R(1)}\right)^{\alpha} \tag{25}$$

we have

$$k_1 + k_2 = 1 \quad \text{and} \quad a_n = \nu\left(\frac{1}{n}\right). \tag{26}$$

A stopping rule derived from (24) requires the estimation of α and the norming constants a_n. For the estimation of the shape parameter α see de Haan [6] or Dorea [4]. The following theorem will provide equivalent results and avoids the estimation of the a_n's.

THEOREM 1. *Under Condition* 1 *we have, for* $y>0$,

$$\lim_{n\to\infty} P\left(-\frac{k_1 y}{n} \leqq X_n - \mathbf{x_0} \leqq \frac{k_2 y}{n}\right) = 1 - \exp\left(-\frac{y}{\mathbf{m}(\Omega)}\right) \tag{27}$$

and

$$\lim_{n\to\infty} P\left(\frac{Y_n - \mathbf{y_0}}{a_n} \leqq y \,\middle|\, -\frac{k_1 y^{\alpha}}{n} \leqq X_n - \mathbf{x_0} \leqq \frac{k_2 y^{\alpha}}{n}\right) = 1 \tag{28}$$

where k_1 *and* k_2 *are defined by* (25).

Before proving the theorem, we will justify the use of our stopping rule 2. Assume that n is large and $n \geqq -(m(\Omega)\log\beta/\varepsilon)$. Then from (27) we have

$$\begin{aligned} P(|X_n - \mathbf{x_0}| \leqq \varepsilon) &\geqq P\left(-\frac{k_1 n\varepsilon}{n} \leqq X_n - \mathbf{x_0} \leqq \frac{k_2 n\varepsilon}{n}\right) \\ &\approx 1 - \exp\left(-\frac{n\varepsilon}{\mathbf{m}(\Omega)}\right) \geqq 1-\beta. \end{aligned}$$

Moreover, by taking $a_n = (1/n)^{1/\alpha}$ and using (28) we have

$$\begin{aligned} P(-\varepsilon^{1/\alpha} \leqq Y_n - \mathbf{y_0} \leqq \varepsilon^{1/\alpha}) &= P\left(-\frac{a_n \varepsilon^{1/\alpha}}{a_n} \leqq Y_n - \mathbf{y_0} \leqq \frac{a_n \varepsilon^{1/\alpha}}{a_n}\right) \\ &= P(-a_n(n\varepsilon)^{1/\alpha} \leqq Y_n - \mathbf{y_0} \leqq a_n(n\varepsilon)^{1/\alpha}) \\ &\approx P(-k_1\varepsilon \leqq X_n - \mathbf{x_0} \leqq k_2\varepsilon) \geqq 1-\beta. \end{aligned} \tag{29}$$

In Dorea [5, p. 46] we can find conditions that enable us to take $a_n = (1/n)^{1/\alpha}$.

Proof. We will borrow the following result from Dorea [5]: for $y>0$ let

$$B_n(y)=\left\{x: f\left(\mathbf{x}_0+\frac{x}{n}\right)\leqq \mathbf{y}_0+a_n y, \left(\mathbf{x}_0+\frac{x}{n}\right)\in\Omega\right\}.$$

Then we have

$$\overline{\lim_{n\to\infty} B_n(y)}=A(y)=[-k_1y^\alpha, k_2y^\alpha] \tag{30}$$

and

$$\lim_{n\to\infty}\mathbf{m}(B_n(y))=y^\alpha, \tag{31}$$

where $\bar{D}$ denotes the closure of the set D.

To prove (27), note that $(X_n=\xi_j)=(f(\xi_j)<\min\{f(\xi_l), 1\leqq l\leqq n, l\neq j\})$ and $(X_n=\xi_j)\cap(X_n=\xi_i)=\phi$ for $i\neq j$. Since $\xi_1,\cdots,\varepsilon_n$ are independent and identically distributed random variables and uniformly distributed on Ω, we have for $E=[-k_1y, k_2y]$

$$\begin{aligned}P(n(X_n-\mathbf{x}_0)\in E)&=\sum_{j=1}^{n}P(n(X_n-\mathbf{x}_0)\in E, X_n=\xi_j)\\&=nP(n(\xi_1-\mathbf{x}_0)\in E, X_n=\xi_1)=nP(n(\xi_1-\mathbf{x}_0)\\&\in E, f(\xi_1)<\min\{f(\xi_2),\cdots,f(\xi_n)\})\\&=\frac{1}{(\mathbf{m}(\Omega))^n}\int_{\{n(u-\mathbf{x}_0)\in E\}}\left(\int_{\{x:x\in\Omega, f(x)>f(u)\}}dx\right)^{n-1}du\\&=\frac{1}{(\mathbf{m}(\Omega))^n}\int_E\left(\int_{\{x:x\in\Omega, f(x)>f(\mathbf{x}_0+r/n)\}}dx\right)^{n-1}dr.\end{aligned}$$

Let $z_n=(f(\mathbf{x}_0+r/n)-\mathbf{y}_0/a_n)$. We can write $(1/\mathbf{m}(\Omega))\int_{\{x\in\Omega, f(x)>f(\mathbf{x}_0+r/n)\}}dx=1-(\mathbf{m}(B_n(z_n))/n\mathbf{m}(\Omega))$ and $P(n(X_n-\mathbf{x}_0)\in E)=$

$$\frac{1}{\mathbf{m}(\Omega)}\left[\int_0^{k_2y}\left(1-\frac{\mathbf{m}(B_n(z_n))}{n\mathbf{m}(\Omega)}\right)^{n-1}dr+\int_{-k_1y}^{0}\left(1-\frac{\mathbf{m}(B_n(z_n))}{n\mathbf{m}(\Omega)}\right)^{n-1}dr\right].$$

From Condition 1 and (26) we have $\lim_{n\to\infty}z_n=R(r)$ where $R(r)=(r/k_2)^{1/\alpha}$ if $r>0$ and $R(r)=(|r|/k_1)^{1/\alpha}$ if $r<0$.

This together with (31) gives

$$\begin{aligned}\lim_{n\to\infty}\left(1-\frac{\mathbf{m}(B_n(z_n))}{n\mathbf{m}(\Omega)}\right)^{n-1}&=\exp\left(-\frac{r}{k_2\mathbf{m}(\Omega)}\right), \qquad r>0\\&=\exp\left(-\frac{|r|}{k_1\mathbf{m}(\Omega)}\right), \qquad r<0.\end{aligned}$$

Finally, an application of the bounded convergence theorem gives

$$\lim_{n\to\infty}P(n(X_n-\mathbf{x}_0)\in E)=(k_1+k_2)\left(1-\exp\left(-\frac{y}{\mathbf{m}(\Omega)}\right)\right).$$

Then (27) follows since $k_1+k_2=1$.

The proof of (28) makes use of essentially the same types of arguments. The fact that

$$\begin{aligned}&P\left(\frac{Y_n-\mathbf{y}_0}{a_n}\leqq y, -k_1y^\alpha\leqq n(X_n-\mathbf{x}_0)\leqq k_2y^\alpha\right)\\&\quad=\frac{1}{\mathbf{m}(\Omega)}\int_{-k_1y^\alpha}^{k_2y^\alpha}\left(1-\frac{\mathbf{m}(B_n(z_n))}{n\mathbf{m}(\Omega)}\right)^{n-1}I_{\{r:f(\mathbf{x}_0+r/n)\leqq\mathbf{y}_0+a_ny\}}\,dr\end{aligned}$$

together with (27) and (30) gives the desired results (I: indicator function). □

Remarks. (a) If $\mathbf{x}_0$ is a right-endpoint, then we require Condition 1 (b) to be satisfied for $z<0$ and we have $k_2=0$. A left-endpoint is similarly treated. The following example illustrates the roles played by α, k_1, and k_2. Let $\mathbf{x}_0\in(0,1)=\Omega$ and $f(x)=3(x-\mathbf{x}_0)^{1/2}$ for $x>\mathbf{x}_0$ and $f(x)=4(|x-\mathbf{x}_0|)^{1/2}$ for $x<\mathbf{x}_0$. Then Condition 1 is satisfied with $\alpha=2$ and $\nu(t)=\sqrt{t}$. It is also satisfied with $\nu(t)=12/6\sqrt{t}$. In this case we have $k_1=9/25$ and $k_2=16/25$, with $k_1+k_2=1$.

(b) If Condition 1 is satisfied with $\Omega\subset R^d$ we shall interpret $\mathbf{x}_0+tz=(\mathbf{x}_0(1)+tz(1),\cdots,\mathbf{x}_0(d)+tz(d))$ and $z\neq(0,\cdots,0)$. In this case for

$$B_n^d(y)=\left\{z: f\left(\mathbf{x}_0+\frac{z}{n^{1/d}}\right)\leqq a_n y\right\}$$

we have

$$\overline{\lim_{n\to\infty}} B_n^d(y)=A^d(y)$$

and

$$\lim_{n\to\infty}\mathbf{m}(B_n^d(y))=\mathbf{m}(A^d(1))y^{d\alpha}.$$

We can choose the function $\nu(t)$ so that $\mathbf{m}(A^d(1))=1$ and (27) and (28) of Theorem 1 become

$$\lim_{n\to\infty} P(n^{1/d}(X_n-\mathbf{x}_0)\in A^d(y^{1/\alpha}))=1-\exp\left(-\frac{y^d}{\mathbf{m}(\Omega)}\right)$$

and

$$\lim_{n\to\infty} P(Y_n-\mathbf{y}_0\leqq a_n y | n^{1/d}(X_n-\mathbf{x}_0)\in A^d(y))=1.$$

Note that $A^d(\varepsilon^{1/\alpha})$ is a neighborhood of zero and $\mathbf{m}(A^d(\varepsilon^{1/\alpha}))=\varepsilon^d$. Assuming that $A^d(\varepsilon^{1/\alpha})\subset[-\varepsilon,\varepsilon]^d$ we can write

$$P((X_n-\mathbf{x}_0)\in[-\varepsilon,\varepsilon]^d)\geqq P(n^{1/d}(X_n-\mathbf{x}_0)\in n^{1/d}A^d(\varepsilon^{1/\alpha}))\approx 1-\exp\left(-\frac{n\varepsilon^d}{\mathbf{m}(\Omega)}\right).$$

The stopping rule 2 then becomes: terminate for n such that,

$$n\geqq-\frac{\mathbf{m}(\Omega)\log\beta}{\varepsilon^d}.$$

3. Concluding remarks. Both stopping rules 1a and 1b require the evaluation of the lower bound

$$\delta_n(\varepsilon)=\frac{\log\beta}{\log(1-\rho_n(\varepsilon)/n)}.$$

The advantage of rule 1b relies on the fact that if at step n we evaluate $\delta_n(\varepsilon)$ then it is enough to verify whether the m repetitions of Y_n satisfy rule 1b. Clearly, if m satisfies rule 1b then we are at step $n+m$ of the algorithm and $n+m$ satisfies rule 1a.

A comparison between rules 1 and 2 can only take place if there exists a unique minimum point $\mathbf{x}_0$ for which the global minimum $\mathbf{y}_0$ is attained. Moreover the shape parameter α of the limiting distribution used for rule 2 needs to be estimated a priori. In fact, from Proposition 1, we have,

$$P(|Y_n-\mathbf{y}_0|\leqq\varepsilon)<1-\beta \quad \text{if } n\geqq\delta_n(\varepsilon),$$

and from Theorem 1 and (29) we have for large n

$$P(|Y_n - \mathbf{y}_0| \leqq \varepsilon) \approx 1 - \exp\left(\frac{n\varepsilon^\alpha}{\mathbf{m}(\Omega)}\right).$$

Hence,

$$P(|Y_n - \mathbf{y}_0| \leqq \varepsilon) \approx 1 - \beta \quad \text{if } n \geqq -\frac{\mathbf{m}(\Omega)\log\beta}{\varepsilon^\alpha} = \delta^*(\varepsilon).$$

Some modest numerical simulations were performed to compare the proposed rules. For the uniform distribution on $\Omega = [0, 1], f(x) = x$ and with several values of ε and β, our numerical results show that:

	β: 0.025	0.05	0.10
ε: 0.01	(1001,242,369)	(985,201,299)	(985,150,230)
0.02	(985,160,184)	(132,132,150)	(132,132,115)
0.05	(985,160,—)	(117,115,—)	(117,98,—)
0.10	(117,117,—)	(97,95,—)	(97,—,—)

where triple on the table represents the stopping step when rule 1a, 1b, or 2 is applied (with $\delta^*(\varepsilon)$ used in place (12)). The unfilled values were those stopping steps for which the algorithm did not provide a corresponding value for Y_n within ε of the global minimum $\mathbf{y}_0$. Similar results were obtained for $f(x) = x^2$ or $\sqrt{x}$ (in this case we have $\alpha = \frac{1}{2}$ or 2, respectively). For $\alpha = \frac{1}{2}$ the performance of rule 2 was considerably worse than for $\alpha = 1$. And for $\alpha = 2$ it was considerably better.

We will now discuss how these rules can be adapted for Algorithm B. Let $\sigma_0 = 0$ and $\sigma_j = \inf\{k;\ k > \sigma_{j-1},\ \xi_k \in \Omega\}$ for $j \geqq 1$. Then, for $S_k = \xi_{\sigma_k}$ and $R_k = f(S_k)$, we have $S_1, S_2, \cdots$ i.i.d. with distribution $G_S(x) = (1/P(\Omega))\int_{\{u:u\in\Omega, u\leqq x\}} dG$ and $R_1, R_2, \cdots$ i.i.d. with distribution $F_R(x) = 1/P(\Omega)\int_{\{u:u\in\Omega, f(u)\leqq x\}} dG$ (see Dorea [4]). Hence, all the previous results can be applied to the subsequence $Y_{\sigma_1}, Y_{\sigma_2}, \cdots$. The stopping rules then become: terminate at step σ_n where n satisfies (6), with $\rho_{\sigma_n}(\varepsilon)$ in place of $\rho_n(\varepsilon)$; or terminate when a value of Y_{σ_n} has been repeated until the step σ_{n+m} where m satisfies (7) with $\rho_{\sigma_n}(\varepsilon)$ in place of $\rho_n(\varepsilon)$.

Now, we can adapt the Algorithm B taking into account the role of the σ_k's by introducing the random variable $Z_0, Z_1, \cdots$ as follows:

step 0. Let $Z_0 = 0$.

step $k+1$. Let $Z_{k+1} = Z_k + 1$ if $\xi_{k+1} \in \Omega$ and $Z_{k+1} = Z_k$, otherwise. The stopping rules then can be rephrased as:

(1a) Terminate the algorithm at step n if

$$Z_n \geqq \frac{\log\beta}{\log\left(1 - \dfrac{\rho_n(\varepsilon)}{Z_n}\right)}.$$

(1b) Terminate the algorithm if for some m and n we have,

$$Z_{n+m} \geqq \frac{\log\beta}{\log\left(1 - \dfrac{\rho_n(\varepsilon)}{Z_n}\right)}.$$

(2) Terminate the algorithm at step n if

$$Z_n \geqq -\frac{\mathbf{m}(\Omega)\log\beta}{\varepsilon}.$$

Acknowledgment. We are grateful to C. R. Gonçalves for the help in the numerical simulations.

REFERENCES

[1] C. G. E. BOENDER, A. H. G. RINNOOY KAN, L. STOUGIE, AND G. T. TIMMER, *A stochastic method for global optimization*, Math. Programming, 22 (1982), pp. 125–140.

[2] H. A. DAVID, *Order Statistics*, John Wiley, New York, 1981.

[3] L. DEVROYE, *Progressive random search of continuous function*, Math. Programming, 15 (1978), pp. 330–342.

[4] C. C. Y. DOREA, *Limiting distribution for random optimizatiom methods*, SIAM J. Control Optim., 24 (1986), pp. 76–82.

[5] ———, *Estimation of the extreme value and the extreme points*, Ann. Inst. Statist. Math., 39 (1987), pp. 37–48.

[6] L. de Haan, *Estimation of the minimum of a function using order statistics*, J. Amer. Statist. Assoc., Th. and Meth., 79 (1981), pp. 467–469.

[7] A. H. G. RINNOOY KAN AND G. T. TIMMER, *Stochastic global optimization methods. Part I: clustering methods*, Math. Programming, 39 (1987), pp. 27–56.

[8] ———, *Stochastic global optimization methods. Part II: multi level methods*, Math. Programming, 39 (1987), pp. 57–78.

[9] F. J. SOLIS AND R. J-B WETS, *Minimization by random search techniques*, Math. Oper. Res., 6 (1981), pp. 19–30.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 851-902, July 1990

006

ON THE EXISTENCE OF OPTIMAL CONTROLS*

U. G. HAUSSMANN† AND J. P. LEPELTIER‡

Abstract. The optimal control problem where the state is governed by an Itô stochastic differential equation (possibly just an ordinary differential equation) is formulated in martingale terms. Under a coercivity condition (which is weaker than compactness of the control set), a convexity condition, and mild continuity hypotheses on the data, it is shown by the direct method that optimal controls exist. Hard and soft constraints are allowed. In the absence of soft constraints it is shown that there exists an optimal control that is a function only of the present time and state, i.e., the synthesis problem has a solution. The main tool here is Krylov's Markovian Selection Theorem.

Key words. existence theory, controlled diffusion, martingale problem, relaxed controls, Markovian selection, synthesis problem

AMS(MOS) subject classifications. 49A60, 49A10, 93B50, 93E20

1. Introduction. There are two general approaches available to establish the existence of optimal controls; either the sufficient conditions of the Hamilton-Jacobi theory are guaranteed, or it is shown that a minimizing sequence of controls is compact (the direct method). This situation prevails not only in the deterministic case but also in the stochastic case where the state satisfies an Itô equation (which may be an ordinary differential equation). In the stochastic control literature early examples of the first approach are the articles by Davis [5] and Bismut [3], and of the direct method articles by Benes [1] and Kushner [15]. All of the work in the literature with the exception of the recent work of Loewen [16] requires the control set to be compact. In [16] this compactness condition is weakened to a coercivity condition—a result well known in the deterministic theory (cf. the book by Fleming and Rishel [10]). Our first result is very similar to Loewen's, but we require a bit less regularity, we allow the diffusion coefficient to depend on the control, we allow hard constraints (i.e., state constraints that must be met almost surely) as well as soft constraints (i.e., constraints that must be met in the mean), and we allow the terminal time to be not merely a fixed time but rather a first exit time or even a stopping time chosen by the controller (optimal stopping). Moreover, the method of proof is quite different from Loewen's; in fact, it hinges on the introduction of relaxed controls, an approach used previously in the stochastic setting by Fleming and Nisio [9] and others (cf. El Karoui, Huu Nguyen, and Jeanblanc-Picqué [8] for further references). A brief survey of the use of relaxed controls is given by Borkar [4].

Recently, Haussmann [11] and El Karoui, Huu Nguyen, and Jeanblanc-Picqué [8] have shown that if the data is bounded and the control set is compact, then in the absence of constraints an optimal control can be found that is a function only of the present time and state, i.e., the synthesis problem can be solved. Our second result is an extension of this result to the case of unbounded data with noncompact control set. The main tool used is Krylov's Markovian Selection Theorem, which enables us to apply an abstract version of dynamic programming. In fact, given our first result

* Received by the editors July 6, 1988; accepted for publication (in revised form) September 8, 1989. This work was supported by National Science and Engineering Research Council grant A8051.

† Mathematics Department, University of British Columbia, 121 1984 Mathematics Road, Vancouver, V6T 1Y4.

‡ Département de Mathématiques et Informatique, Université du Maine, Route de Laval, 72017 LeMans, France.

(the existence theorem) the proof is similar in spirit to the one given for the bounded-compact case, but new technical difficulties do arise.

We should mention that similar results were obtained by El Karoui [7] under the severe restriction (among others) that the diffusion coefficient be nonsingular. Our method is quite different, but in § 5 we have borrowed the idea that the "mixed" control problem can be solved by first solving an optimal stopping problem and then a control problem where the stopping time is not controlled.

We state the problem precisely and make some observations in § 2. In § 3 we reformulate it as a martingale problem and we introduce relaxed controls and canonic relaxed controls that are called control rules. We do this at some length for the benefit of the uninitiated reader since the level of complexity is nontrivial: we are dealing with measures defined on sets of measures. In addition we show that under our hypotheses existence of an optimal control in any of these forms guarantees existence of an optimal control in any other form. Then in § 4 we prove the existence of an optimal control rule, and in § 4.10 we interpret this result in the deterministic case for the reader who wishes to avoid probability theory. In § 5 we establish that the synthesis problem has a solution and briefly mention an example. The Appendix contains some technical lemmas.

2. The control problem. To formulate the control problems we require some notation, which we collect here.

- $\mathbb{R}_+ = [0, \infty)$ and $\bar{\mathbb{R}}_+ = [0, \infty]$. Similarly, $\mathbb{R}_+^m$, $\bar{\mathbb{R}}_+^m$ are the m-dimensional analogues of $\mathbb{R}_+$, $\bar{\mathbb{R}}_+$. $\bar{\mathbb{R}}^d = [-\infty, \infty]^d$. If $D \subset \mathbb{R}_+ \times \mathbb{R}^d$, then $\bar{D}$ is the closure of D and

$$\bar{D}_\infty = \{(\infty, x) \in \bar{\mathbb{R}}_+ \times \bar{\mathbb{R}}^d : \text{there exist } x_n \to x,\ t_n \to \infty,\ (t_n, x_n) \in D\},$$

$$\bar{\bar{D}} = \bar{D} \cup \bar{D}_\infty.$$

- $C(\mathbb{R}_+; \mathbb{R}_+^d)$ is the space of continuous functions from $\mathbb{R}_+$ into $\mathbb{R}^d$ with the topology of uniform convergence on compact intervals. It will be abbreviated to C.
- $C_b(\mathbb{R}^d)$ is the set of bounded continuous functions from $\mathbb{R}^d$ into $\mathbb{R}$.
- $C_b^2(\mathbb{R}^d)$ is the set of functions in $C_b(\mathbb{R}^d)$ that have two bounded continuous derivatives.
- $C_c^\infty(\mathbb{R}^d)$ is the set of functions in $C_b(\mathbb{R}^d)$ that are infinitely differentiable and have compact support.
- $\mathbb{S}^d$ is the set of symmetric $d \times d$ matrices, $\mathcal{S}^d$ will denote its Borel σ-algebra.
- If A is a metric space, then $\mathcal{A}$ will denote its Borel σ-algebra.
- If X is a random variable on $(\Omega, \mathcal{F}, P)$, then the expectation of X is denoted by $P(X)$.
- For x in $C(\mathbb{R}_+; \mathbb{R}^d)$

$$\|x\|_t = \sup\{|x_s| : 0 \leqq s \leqq t\}.$$

- For a and b in $\mathbb{R}$, $a \wedge b = \min\{a, b\}$.
- If A is a separable metric space, then $M_+(A)$ (respectively, $M_1(A)$) denotes the bounded nonnegative Radon (respectively, probability) measures on $(A, \mathcal{A})$ with the topology of weak convergence. Both $M_+(A)$ and $M_1(A)$ are separable metric spaces, and are complete if A is (cf. [6, III, 60]). Furthermore, comp $(M_1(A))$ denotes the metric space of nonempty compact subsets of $M_1(A)$ under the Hausdorff metric.
- If $R : S \to \text{comp}(A)$, i.e., R is a multifunction from S into A, both of which are metric spaces, then meas (R) is the set of measurable selections of R, i.e., y is in meas (R) if $y : S \to A$ is Borel measurable and $y(s)$ lies in $R(s)$ for each s. If R is measurable, then meas (R) is not empty (cf. [19, Thm. 12.1.10]).

2.1. The data. We are given the following data.

- U, a closed subset of the Euclidean space (in fact, a closed, σ-compact subset of a Banach space would do, as would a Polish space if $p=0$; cf. below).
- D, a subset of $\mathbb{R}_+\times\mathbb{R}^d$, open in the relative topology of $\mathbb{R}_+\times\mathbb{R}^d$.
- (a, b): $D\times U\to\mathbb{S}^d\times\mathbb{R}^d$, measurable, such that

$$(x, u)\to(a(t, x, u), b(t, x, u))$$

is continuous for each t, and such that there exist nonnegative constants k, β, γ, ν, p with $0\leqq\beta\leqq 2$, $\nu\leqq p$, $\gamma\bar{\beta}\leqq p$, where $\bar{\beta}=\max\{1,\beta\}$, for which

$$\begin{aligned}|a(t, x, u)|&\leqq k(1+|x|^{\beta}+|u|^{\nu}),\\ |b(t, x, u)|&\leqq k(1+|x|+|u|^{\gamma});\end{aligned}\tag{2.1}$$

we call p the exponent of coercivity (cf. (3.5)).

- $f: D\times U\to\bar{\mathbb{R}}_+^{1+m}$, measurable, such that each component of f is lower semicontinuous in (x, u) for each t.
- g: $\bar{D}\to\mathbb{R}_+^n$, continuous, constant on $\bar{D}_\infty$.
- h: $\bar{D}\to\bar{\mathbb{R}}_+^{1+m}$, such that each component of h is lower semicontinuous, constant on $\bar{D}_\infty$.
- (λ^1, λ^2), an element of $\mathbb{R}^m\times\mathbb{R}^n$.

We observe that in the deterministic case (cf. [10, Chap. III]), we may take $p=\gamma=1$, $\beta=\nu=0$, and in Loewen's case [16] $p\geqq 2$, $\gamma=1$, $\beta\leqq 2$ (although he also requires that $\beta<p$ and that a be independent of u, i.e., $\nu=0$).

DEFINITION 2.2. Given an initial condition (s, x) in D we say that a (strict) *control* is a term

$$\alpha=(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}_{t\geqq s}, \{X_t\}_{t\geqq s}, \{u_t\}_{t\geqq s}, S)$$

where

(C_1) $(\Omega, \mathcal{F}, P)$ is a probability space with filtration $\{\mathcal{F}_t\}$.

(C_2) $\{u_t\}$ is a U-valued, $\{\mathcal{F}_t\}$ progressively measurable process such that for each $T\geqq s$

$$P\left(\int_s^{T\wedge\rho\wedge S}|u_t|^p\,dt\right)<\infty,$$

where ρ is the first exit time of (t, X_t) from D, i.e., $\rho(\omega)=\inf\{t\geqq s: (t, X_t(\omega))\notin D\}$.

(C_3) $\{X_t\}$ is an $\mathbb{R}^d$-valued right continuous, almost surely (a.s.) continuous, progressively measurable process such that for some pair $(\sigma(t, x, u), \{w_t\}_{t\geqq s})$, $\{X_t\}$ satisfies

$$X_t=x+\int_s^t b(\theta, X_\theta, u_\theta)\,d\theta+\int_s^t\sigma(\theta, X_\theta, u_\theta)\,dw_\theta,\qquad s\leqq t\leqq\rho\wedge S,\quad\text{a.s.}\tag{2.2}$$

where $\sigma(t, x, u)$ is a $(d\times d')$-dimensional matrix with $\sigma(t, x, u)\sigma(t, x, u)'=a(t, x, u)$ ($'$ denotes transpose) and where $\{w_t\}_{t\geqq s}$ is a standard Brownian motion on $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$.

(C_4) S is a stopping time, i.e., a measurable function $\Omega\to\bar{\mathbb{R}}_+$ such that

$$\{\omega: S(\omega)\leqq t\}\in\mathcal{F}_t,\qquad S\geqq s\quad\text{a.s.}$$

We write $\mathcal{U}_{sx}$ for the set of all such controls. For α in $\mathcal{U}_{sx}$, $\{u_t\}$ is called the control process, $\{X_t\}$ the state process, and S the stopping time. $\mathcal{F}_t$ specifies the history of the world up to time t. We write $\mathcal{F}_t^X$ for the history of $\{X_t\}$ to time t, i.e., $\mathcal{F}_t^X \subset \mathcal{F}_t$ is the σ-algebra generated by the family of random variables $\{X_\theta : s \leqq \theta \leqq t\}$. Let us define X_t on $\mathbb{R}_+$ by setting $X_t = x$ for $0 \leqq t \leqq s$.

Let $y = \rho \wedge S$ and $X_t^* = X_{t\wedge y}$. Note that y is a stopping time. The following lemma shows that the integrals in (2.2) make sense, and gives a bound on the moment of order $\bar{\beta}$ of $\|X^*\|_t$, which among other things implies that the stochastic integral in (2.2) is a martingale. Let $\bar{p} = p \min\{\gamma^{-1}, 2\nu^{-1}\}$.

LEMMA 2.3. *For α in $\mathcal{U}_{sx}$ the integrals in* (2.2) *are well defined and for q in $[\bar{\beta}, \bar{p}]$ there exists a constant k_q depending only on q such that*

$$P(\|X^*\|_t^q) \leqq k_q \exp(k_q(t-s)^q)\left\{1 + |x|^q + (t-s)^q + P\int_s^{t\wedge y} (|u_\theta|^p)\, d\theta\right\}. \tag{2.3}$$

Proof. The first statement follows from (2.3) with $q = \bar{\beta}$. To establish (2.3) we define

$$\tau_N = \inf\{t \geqq s : |X_t| \geqq N\} \wedge y$$

and set $X_t^N = X_{t\wedge\tau_N}$. Then from (2.2)

$$X_t^N = x + \int_s^{t\wedge\tau_N} b(\theta, X_\theta^N, u_\theta)\, d\theta + \int_s^{t\wedge\tau_N} \sigma(\theta, X_\theta^N, u_\theta)\, dw_\theta$$

since now the integrals are well defined according to (C_2) and (2.1). Now for suitable constants $\bar{k}_q$ and $\tilde{k}_q$ the Burkholder–Davis–Gundy inequality gives

$$\begin{aligned} P(\|X^N\|_t^q) &\leqq \bar{k}_q\left\{|x|^q + (t-s)^{q-1} P\int_s^{t\wedge y} [1 + \|X^N\|_\theta^q + |u_\theta|^{\gamma_q}]\, d\theta \right. \\ &\qquad \left. \cdot P\left[\left(\int_s^{t\wedge y} |a(\theta, X_\theta^N, u_\theta)|\, d\theta\right)^{q/2}\right]\right\} \\ &\leqq \tilde{k}_q\left\{|x|^q + 1 + (t-s)^q + [(t-s)^{q-1} + 1]\right. \\ &\qquad \left. \cdot \left[\int_s^t P(\|X^N\|_\theta^q)\, d\theta + P\int_s^{t\wedge y} |u_\theta|^p\, d\theta\right]\right\}. \end{aligned} \tag{2.4}$$

We have used the fact that for any random variable y

$$P(|y|^{q/2}) \leqq 1 + P(|y|)$$

if $q/2 \leqq 1$. Now (2.3) follows for X^N by applying Gronwall's inequality to (2.4). Since $\|X^N\|_t = \|X^*\|_t \wedge N$ then (2.3) follows by the Monotone Convergence Theorem. □

Note that if $\nu = \gamma = 0$ (for example, U compact) then $\bar{p} = +\infty$ and (2.3) holds for $q \geqq \bar{\beta}$. Moreover, if a and b are bounded, then (2.3) holds without the term involving u and without the exponential factor.

Let us now define the cost and constraints. For α in $\mathcal{U}_{sx}$ we set

$$\begin{aligned} F(\alpha, \omega) &= \int_s^y f(t, X_t, u_t)\, dt + h(y, X_y), \\ G(\alpha, \omega) &= g(y, X_y) \end{aligned} \tag{2.5}$$

and we denote the components of F by F_i, $i = 0, 1, \cdots, m$, and those of G by G_i, $i = 1, 2, \cdots, n$. Note that $\int f_i\, dt$ is well defined, possibly $+\infty$, since $f_i \geqq 0$. $g(y, X_y)$ and $h(y, X_y)$ are also well defined even if $y = +\infty$ because we assume g and h to be constant

on $\bar{D}_\infty$. The usual case of discounting would imply that this constant value is zero. We say that α in $\mathcal{U}_{sx}$ is *feasible* if

$$J_0^1(s,\alpha) := P(F_0(\alpha,\cdot)) < \infty,$$
$$J_i^1(s,\alpha) := P(F_i(\alpha,\cdot)) \leqq \lambda_i^1, \qquad i = 1, 2, \cdots, m,$$
$$J_i^2(s,\alpha) := P(G_i(\alpha,\cdot)) = \lambda_i^2, \qquad i = 1, 2, \cdots, n.$$

We write $\mathcal{U}_{sx}^f$ for the set of feasible controls in $\mathcal{U}_{sx}$; this set may be empty (it certainly is if $\lambda_i^j < 0$ for some i, j). Let us also set $J = (J^1, J^2)$.

The strict control problem can now be stated precisely:

$$\inf\{J_0^1(s,\alpha): \alpha \in \mathcal{U}_{sx}^f\}. \tag{2.6}$$

The problem that concerns us is the *existence* of a minimizing control. We are only concerned with the (random) time interval $s \leqq t \leqq y(\omega)$ so we can redefine X_t and u_t arbitrarily for $t > y(\omega)$ (cf. § 4). We have already set $X_t = x$ on $0 \leqq t < s$; now let $u^0 \in U$ be a fixed but arbitrary element and set $u_t(\omega) = u^0$ on $0 \leqq t < s$.

Remark 2.4. Here we add a few observations concerning the form of the problem. First, concerning the constraints, the conditions $h_i \geqq 0, f_i \geqq 0$ can be relaxed to $h_i(t,x) \geqq -M > -\infty$, $f_i(t,x,u) \geqq -\bar{f}_i(t)$ where $\bar{f}_i \geqq 0$ and integrable on $[0,\infty)$. Now set

$$\tilde{h}_i(t,x) = h_i(t,x) + M + \int_t^\infty \bar{f}_i(\theta)\, d\theta,$$
$$\tilde{f}_i(t,x,u) = f_i(t,x,u) + \bar{f}_i(t),$$
$$\tilde{\lambda}_i = \lambda_i + M + \int_s^\infty \bar{f}_i(t)\, dt.$$

Then $\tilde{h}_i \geqq 0$, $\tilde{f}_i \geqq 0$ and these two functions have the same measurability and semicontinuity as h_i, f_i. Moreover, $J_i^1 \geqq \lambda_i$ if and only if

$$\tilde{J}_i^1 := P\left(\int_s^{S\wedge\rho} \tilde{f}_i\, dt + \tilde{h}_i(\rho\wedge S, X(\rho\wedge S))\right) \leqq \tilde{\lambda}_i.$$

Allowing f_0 and h_0 to assume the value $+\infty$ permits us to introduce *hard* constraints, i.e., constraints that must hold almost surely as opposed to the soft constraints $J_i^1(s,\alpha) \leqq \lambda_i^1$, $J_i^2(s,\alpha) = \lambda_i^2$, $i > 0$. Indeed if $f_0(t,x,u) = +\infty$ for (t,x) in $A \subset D$, A open, and if $h_0(t,x) = +\infty$ for (t,x) in $B \cap \bar{D}$, B open, then α is feasible only if (y, X_y) is almost surely not in B and for each $t < y$, X_t is almost surely not in A_t, the t-section of A.

We have chosen to give the exposition for the case when the data depend on the state X only through its present value X_t. In fact, the results of this paper excluding § 5 go through if the data are allowed to depend on the past of X. The main change is to replace $|x|$ in (2.1) by $\|x\|_t$ for x in $C(\mathbb{R}_+;\mathbb{R}^d)$.

Remark 2.5. We are allowing a great deal of latitude in the notion of control; in particular, since $(\Omega, \mathcal{F}, P)$ and $\{w_t\}$ are not specified a priori, then $\{X_t\}$ is a weak solution of (2.2).

Leaving $(\sigma, \{w_t\})$ ambiguous in the definition of α in $\mathcal{U}_{sx}$ is quite appropriate. In fact, $\{w_t\}$ is usually not observed; we only know that it is a Brownian motion, so it makes good sense to leave it unspecified. Moreover, as is well known, a, and not σ, is intrinsic to the process $\{X_t\}$, so we have taken it as the given.

Nevertheless we would like to formulate the problem without reference to the ambiguous $(\sigma, \{w_t\})$. This we can do by reformulating the problem in terms of the solution of a martingale problem (cf. Proposition 3.1). Beyond this we would also like

to restrict $(\Omega, \mathcal{F}, \{\mathcal{F}_t\})$ to something physically realizable. In Corollary 3.9 we show that if S is an $\{\mathcal{F}_t^X\}$ stopping time then we may restrict Ω to be the canonic space $C(\mathbb{R}_+; \mathbb{R}^d)$ (the trajectories of $X_\cdot$) so that we are using the *natural controls* (in the terminology of Krylov [14]), i.e., control processes of the form

$$u_t(\omega) = v(t, X_\cdot(\omega))$$

with

$$v: \mathbb{R}_+ \times C(\mathbb{R}_+; \mathbb{R}^d) \to U,$$

and v progressively measurable. This means that $\mathcal{F}_t = \mathcal{F}_t^X$. However we prefer to let Ω be the canonic space of trajectories $(X_\cdot, u_\cdot, S)$ (after introducing relaxed controls). That this is permissible is established in Theorem 3.13. The law of $(X_\cdot, u_\cdot, S)$ is called a control rule.

Remark 2.6. Let us consider two special cases. The first is the typical deterministic problem. If

$$dX_t = b(t, X_t, u_t)\, dt, \qquad X_0 = x,$$

find u and S to minimize

$$\int_0^S f_0(t, X_t, u_t)\, dt + h_0(S, X_S)$$

such that

$$\tilde{g}_i(S, X_S) = \lambda_i^2, \qquad i = 1, \cdots, n.$$

We incorporate the equality constraints into h_0 by redefining $h_0(s, x) = +\infty$ on the complement of

$$\{(s, x): \tilde{g}_i(s, x) = \lambda_i^2, \qquad i = 1, \cdots, n\}.$$

This problem now has the form discussed with $a = 0$, $D = \mathbb{R}_+ \times \mathbb{R}^d$, $m = 0$, $n = 0$ (i.e., $g_i = 0$).

However, we are allowing the controls to be *randomized.* Note that if u is deterministic we may take Ω as a singleton $\{\omega\}$, and P then is a unit mass on $\{\omega\}$. Then

$$\int_0^S f_0\, dt + h_0 = P\left\{\int_0^S f_0\, dt + h_0\right\} \tag{2.7}$$

and hence u generates a feasible control with the same cost. Conversely, suppose that $\tilde{\alpha}$ in $\mathcal{U}_{0x}^f$ is optimal. We may then ask whether it generates a deterministic pair $(\bar{u}, \bar{S})$ which solves the original deterministic problem where the cost is not taken in the average sense. Certainly, if

$$\lambda_0 := \tilde{P}\left\{\int_0^{\tilde{S}} f_0\, dt + h_0\right\},$$

then $\tilde{P}\{\omega: \int_0^{\tilde{S}} f_0\, dt + h_0 \leqq \lambda_0\} \neq 0$, so there exists ω_0 such that

$$\int_0^{\tilde{S}(\omega_0)} f_0(t, \tilde{X}_t(\omega_0), \tilde{u}_t(\omega_0))\, dt + h_0(\tilde{S}(\omega_0), \tilde{X}_{\tilde{S}(\omega_0)}(\omega_0)) \leqq \lambda_0.$$

On the other hand, if (S, u) with corresponding X is feasible for the deterministic problem then by (2.7) and the optimality of $\tilde{\alpha}$ its cost is at least λ_0, and hence no smaller than the cost associated with $(\bar{S}, \bar{u}_\cdot) := (\tilde{S}(\omega_0), \tilde{u}_\cdot(\omega_0))$. But this also means that

the cost associated with $(\bar{S}, \bar{u})$ can be no smaller than λ_0, and hence equals λ_0. Thus under $\tilde{\alpha}$ almost all sample paths must have cost $=\lambda_0$, since λ_0 is the mean cost and no paths can have smaller cost. Hence $\tilde{P}$-a.s. $(\tilde{S}(\omega), \tilde{u}_\cdot(\omega))$ is optimal for the original deterministic problem and the minimal cost is λ_0.

Next let us turn to the optimal stopping problem. If $D=\mathbb{R}_+\times\mathbb{R}^d$ and the data are independent of u, then we have an optimal stopping problem. To fix the process X we require b and σ to be Lipschitz in x so that we have unique strong solutions of (2.1). Of course we admit randomized stopping times S in Definition 2.2 but we will see in § 5 that we may take the optimal S to be an $\mathscr{F}_t^X$ stopping time (hence not randomized).

Let us finally remark that in our control problem, if $h_0=+\infty$ on D then S is never active so $y=\rho$ and the controller cannot choose when to stop.

3. The martingale model. We begin by removing the ambiguous term $(\sigma, \{w_t\})$ from the model and then we introduce canonic controls. For ϕ in $C_b^2(\mathbb{R}^d)$ we define

$$L\phi(t, x, u)=\frac{1}{2}\sum_{ij} a^{ij}(t, x, u)\phi_{x_ix_j}(x)+\sum_i b^i(t, x, u)\phi_{x_i}(x)$$

where $\{a^{ij}\}$ are the entries of a, $\{b^i\}$ are the components of b, $\phi_{x_i}=\partial\phi/\partial x_i$ and $\phi_{x_ix_j}=\partial^2\phi/\partial x_i\,\partial x_j$.

PROPOSITION 3.1. *α is in $\mathscr{U}_{sx}$ if and only if it satisfies* (C_1), (C_2), (C_4), *and the following*:

(C_3') *$\{X_t\}$ is an $\mathbb{R}^d$-valued right-continuous, almost surely continuous, progressive process on $(\Omega, \mathscr{F}, \{\mathscr{F}_t\})$ such that for any ϕ in $C_b^2(\mathbb{R}^d)$, $M_t^*(\phi, \alpha)$ is a $(P, \{\mathscr{F}_t\})$ martingale for $t\geqq s$, where $M_t^*(\phi, \alpha)=M_{t\wedge y}(\phi, \alpha)$ and*

$$M_t(\phi, \alpha):=\phi(X_t)-\int_s^t L\phi(\theta, X_\theta, u_\theta)\,d\theta.$$

Proof. This is the result of Ikeda and Watanabe [12, Prop. 2.1, Chap. IV] (cf. also [19, Thm. 4.5.2]). If a is degenerate, we must enlarge Ω in going from (C_3') to (C_3). □

From now on we will use (C_1), (C_2), (C_3'), (C_4) to define $\mathscr{U}_{sx}$. Let us next introduce the relaxed controls in this setting—we will need them to define the control rules later. Note that $M_1(U)$ is a Polish space since U is one (cf. the notation at the beginning of § 2). If ϕ is a measurable function mapping $U\to\mathbb{R}$ with

$$|\phi(u)|\leqq k(1+|u|^p),$$

then we can extend ϕ to

$$\left\{\mu\in M_1(U)\colon \int_U |u|^p\mu(du)<\infty\right\}$$

by

$$\phi(\mu)=\int_U \phi(u)\mu(du):=\langle\phi, \mu\rangle.$$

DEFINITION 3.2. Given an initial condition (s, x) in D, we say that α is a *relaxed control*, i.e., α in $\tilde{\mathscr{U}}_{sx}$, if

$$\alpha=(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\}, \{X_t\}, \{\mu_t\}, S)$$

where $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$ satisfy (C_1), where S satisfies (C_4), where $\{\mu_t\}$ is a progressively measurable $M_1(U)$-valued process such that for each T in $[s, \infty)$

$$P\left(\int_s^{T\wedge y} |\mu_t|^p \, dt\right) < \infty,$$

and where $\{X_t\}$ is a right-continuous, almost surely continuous progressively measurable $\mathbb{R}^d$-valued process such that $P\{X_s = x\} = 1$ and for any ϕ in $C_b^2(\mathbb{R}^d)$ $M_t^*(\phi, \alpha)$ is a $(P, \{\mathcal{F}_t\})$ martingale for $t \geqq s$, where $M_t^*(\phi, \alpha) = M_{t\wedge y}(\phi, \alpha)$ and

$$M_t(\phi, \alpha) := \phi(X_t) - \int_s^t L\phi(\theta, X_\theta, \mu_\theta) \, d\theta.$$

Note that according to our notation

$$|\mu_t|^p = \int_U |u|^p \mu_t(du),$$

and that $\mu_t(\omega)$ is progressively measurable if for each $t < \infty$, $(\theta, \omega) \to \mathbb{1}_{[0,t]}(\theta)\mu_\theta(\omega)$ is $\mathcal{R}_+ \times \mathcal{F}_t \to \mathcal{M}_1(U)$ measurable.

We observe that the analogue of Proposition 3.1 holds true, i.e., on some extension of $(\Omega, \mathcal{F}, P)$ there exists a pair $(\sigma, \{w_t\})$ such that $\sigma(t, x, \mu)\sigma(t, x, \mu)' = a(t, x, \mu)$ and $\{w_t\}$ is a standard Wiener process such that

$$X_t = x + \int_s^t b(\theta, X_\theta, \mu_\theta) \, d\theta + \int_s^t \sigma(\theta, X_\theta, \mu_\theta) \, dw_\theta \quad \text{a.s.} \tag{3.1}$$

It is possible to relate $\sigma(t, x, \mu)$ to a square root of $a(t, x, u)$ as follows. If $\bar{\sigma}(t, x, u)$ is any square root of $a(t, x, u)$, then one choice of σ is $\sigma(t, x, \mu) = (\bar{\sigma}(t, x, \mu), s(t, x, \mu))$ where s satisfies

$$s(t, x, \mu)s(t, x, \mu)' = a(t, x, \mu) - \bar{\sigma}(t, x, \mu)\bar{\sigma}(t, x, \mu)'$$

(cf. [8, Thm. 2.5]).

It follows from (2.1) that

$$\begin{aligned} |a(t, x, \mu)| &\leqq k(1 + |x|^\beta + |\mu|^\nu), \\ |b(t, x, \mu)| &\leqq k(1 + |x| + |\mu|^\gamma). \end{aligned} \tag{3.2}$$

Now, as in Lemma 2.3, we obtain Lemma 3.3 from (3.1).

LEMMA 3.3. *For q in $[\bar{\beta}, \bar{p}]$ there exists a constant k_q such that for any α in $\tilde{\mathcal{U}}_{sx}$*

$$P[\|X^*\|_t^q] \leqq k_q \exp(k_q(t-s)^q)\left\{1 + |x|^q + (t-s)^q + P\int_s^{t\wedge y} [|\mu_\theta|^p] \, d\theta\right\}.$$

If α is in $\tilde{\mathcal{U}}_{sx}$, then we define G as in (2.5) and

$$F(\alpha, \omega) := \int_s^y \int_U f(t, X_t, u)\mu_t(du) \, dt + h(y, X_y).$$

Now $J_i^j(s, \alpha)$ are defined as previously. We can now define the *feasible relaxed controls* $\tilde{\mathcal{U}}_{sx}^f$ as those α in $\tilde{\mathcal{U}}_{sx}$ for which

$$\begin{aligned} &J_0^1(s, \alpha) < \infty, \\ &J_i^1(s, \alpha) \leqq \lambda_i^1, \qquad i = 1, 2, \cdots, m, \\ &J_i^2(s, \alpha) = \lambda_i^2, \qquad i = 1, 2, \cdots, n. \end{aligned}$$

The relaxed control problem corresponding to the initial condition (s, x) is

$$\inf\{J_0^1(s, \alpha): \alpha \in \tilde{\mathcal{U}}_{sx}^f\}. \tag{3.3}$$

Remark 3.4. It is clear that we can imbed $\mathcal{U}_{sx}^f$ in $\tilde{\mathcal{U}}_{sx}^f$; indeed if α is in $\mathcal{U}_{sx}^f$ with corresponding control process $\{u_t^\alpha\}$, we can set $\mu_t(du) = \delta_t^\alpha(du)$, where $\delta_t^\alpha(\cdot)$ is a unit point mass at u_t^α in U. Hence the inf in (3.3) will be no greater than that in (2.6). The converse, and hence equality, follows under suitable regularity hypotheses (cf. [8, Thm. 4.11]).

Rather than extend this result to the noncompact case we show that under a convexity hypothesis and a coercivity hypothesis each feasible relaxed control (i.e., each element of $\tilde{\mathcal{U}}_{sx}^f$) corresponds to a feasible strict control (i.e., an element of $\mathcal{U}_{sx}^f$) that has a cost no larger than the cost associated with the original relaxed control, and hence the strict problem (2.6) and the relaxed problem (3.3) are equivalent, i.e., a solution of the relaxed problem (3.3) gives rise to a solution of (2.6). Then we will show that (3.3) does indeed have a solution.

For each (t, x) in D we define a set in $\mathcal{S}^d \times \mathcal{R}^d \times \mathcal{R}_+^{1+m}$ by

$$K(t, x) = \{(a(t, x, u), b(t, x, u), z): u \in U, z \in \mathbb{R}^{1+m}, z_i \geqq f_i(t, x, u), i = 0, 1, \cdots, m\}$$

and then we assume that

(3.4) For almost all t and all x such that (t, x) is in D, $K(t, x)$ is convex.

Observe that in $K(t, x)$ we only consider u such that all f_i are finite.

We will also require that $K(t, x)$ be closed, but this is implied by our hypotheses and the following coercivity condition:

(3.5) There exists l in $\{0, 1, \cdots, m\}$ and $\tilde{f}$ in $C(U; \mathbb{R}_+)$ such that for all (t, x, u) in $D \times U$

$$f_l(t, x, u) \geqq \tilde{f}(u), \qquad \lim_{\substack{|u| \to \infty \\ u \in U}} |u|^{-p} \tilde{f}(u) = +\infty.$$

This condition implies that there exists a sequence $\nu_m \to 0$ such that if $|u| > m$ then $|u|^p \leqq \nu_m \tilde{f}(u)$. Of course if U is compact, then (3.5) holds trivially.

PROPOSITION 3.5. *Assume* (3.5). *Then* $K(t, x)$ *is closed.*

Proof. If (a, b, z) is in $K(t, x)$ then there is a point u in U and v in $\mathbb{R}_+^{1+m}$ such that $z = f(t, x, u) + v$. Assume that (dropping the (t, x))

$$(a(u_n), b(u_n), f(u_n) + v_n) \to (a, b, z).$$

Since $f(u_n)$ and v_n are in $\mathbb{R}_+^{1+m}$ then both sequences must be bounded. Hence a subsequence $v_{n_k} \to v_0$. Moreover, the sequence $\{u_n\}$ is bounded. Indeed, according to (3.5), if $|u_n| \to \infty$ then $\tilde{f}(u_n) \to \infty$, i.e., $f_l(t, x, u_n) \to \infty$, contradicting the boundedness of $\{f_l(u_n)\}$. Hence for a further subsequence $u_{n_{k_m}} \to u_0$ and u_0 is in U since U is closed. The continuity of $a(\cdot)$ and $b(\cdot)$ imply that $a = a(u_0)$, $b = b(u_0)$. From the lower semicontinuity of f we obtain $z_i \geqq f_i(u_0) + (v_0)_i$ so $z = f(u_0) + v$ where $v_i = z_i - f_i(u_0) \geqq (v_0)_i \geqq 0$. Thus (a, b, z) is in $K(t, x)$. □

We can now show that subject to (3.4), (3.5), any feasible relaxed control gives rise to a feasible (strict) control without increasing the cost.

THEOREM 3.6. *Assume* (3.4) *and* (3.5). *If* $\tilde{\alpha}$ *is in* $\tilde{\mathcal{U}}_{sx}^f$ *then there exists a control process* $\{u_t\}$ *such that*

$$\alpha := (\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{X_t\}, \{u_t\}, S)$$

is in $\mathcal{U}_{sx}^{f}$ *and*

$$
\begin{aligned}
&J_i^1(s,\alpha) \leqq J_i^1(s,\tilde{\alpha}), \qquad i=0,1,2,\cdots,n,\\
&J_i^2(s,\alpha) = J_i^2(s,\tilde{\alpha}), \qquad i=1,2,\cdots,m.
\end{aligned} \tag{3.6}
$$

Proof. Given $\tilde{\alpha}$ in $\tilde{\mathcal{U}}_{sx}^{f}$,

$$\tilde{\alpha} = \{\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{X_t\}, \{\mu_t\}, S\}$$

we define c by

$$
\begin{aligned}
c(t,\omega) &= (a,b,f)(t, X_t(\omega), \mu_t(\omega))\\
&= \int_U (a,b,f)(t, X_t(\omega), u)\mu_t(\omega, du).
\end{aligned}
$$

Since $K(t,x)$ is closed and convex it follows that $c(t,\omega)$ is in $K(t, X_t(\omega))$ for almost all (t,ω); moreover, c is progressively measurable. Now Theorem A.9 in the Appendix implies that there are progressively measurable processes $\{u_t\}, \{v_t\}$, U, and $\mathbb{R}_+^{1+m}$-valued, respectively, such that for almost all (t,ω)

$$c(t,\omega) = (a,b,f)(t, X_t(\omega), u_t(\omega)) + (0,0,v_t(\omega)).$$

We define $\alpha = (\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{X_t\}, \{u_t\}, S)$. Then $L\phi(t, X_t, \mu_t) = L\phi(t, X_t, u_t)$ except on a (t,ω) null set. Hence for ϕ in $C_c^\infty(\mathbb{R}^d)$ and all $t \geqq s$,

$$M_t(\phi,\tilde{\alpha}) = M_t(\phi,\alpha) \quad \text{a.s.}$$

Now according to Proposition 3.1, α is in $\mathcal{U}_{sx}$ provided for any $T<\infty$

$$P\left(\int_s^{T\wedge y} |u_t|^p\, dt\right) < \infty.$$

But by (3.5) there exist constants m and ν_m such that for any (t,x) in D and any u in U with $|u| \geqq m$

$$|u|^p \leqq \nu_m \tilde{f}(u) \leqq \nu_m f_l(t,x,u).$$

Thus

$$|u|^p \leqq \max\{m^p, \nu_m f_l(t,x,u)\}$$

for all (t,x,u). Hence

$$
\begin{aligned}
P\left(\int_s^{T\wedge y} |u_t|^p\, dt\right) &\leqq Tm^p + \nu_m P\left[\int_s^y f_l(t, X_t, u_t)\, dt\right]\\
&\leqq Tm^p + \nu_m J_l^1(s,\tilde{\alpha})
\end{aligned}
$$

since

$$
\begin{aligned}
f_l(t, X_t, \mu_t) &= f_l(t, X_t, u_t) + (v_t)_l\\
&\geqq f_l(t, X_t, u_t).
\end{aligned}
$$

But $\tilde{\alpha}$ is feasible so $J_l(s,\tilde{\alpha}) < \infty$, and hence α is in $\mathcal{U}_{sx}$. The feasibility of α follows from (3.6), which in turn follows from

$$f_i(t, X_t, \mu_t) = f_i(t, X_t, u_t) + (v_t)_i \geqq f_i(t, X_t, u_t),$$

so we are done. □

Hence under the conditions (3.4), (3.5) we know that the strict problem (2.6) and the relaxed problem (3.3) have the same infimum, and if minimizing controls exist for one problem then they exist for the other.

Observe that if the equality constraint J^2 were to contain an integral term, then we would require the integrand to have the same convexity and regularity as b (rather than f) to make Theorem 3.6 hold. Hence we prefer to change integral constraints of this form to terminal constraints by the addition of another component to the state X, recalling that a need not be nonsingular.

We can now turn to the problem of choosing a "canonic" Ω. The following technical result is useful. Let $\{\mathcal{F}^{X,S}\}$ be the filtration generated by $(X_\cdot, \mathbb{1}_{\{S\leqq\cdot\}})$.

LEMMA 3.7. *If α is in $\tilde{\mathcal{U}}_{sx}$ and if $\{\mathcal{G}_t\}$ is a filtration such that $\mathcal{F}_t^{X,S} \subset \mathcal{G}_t \subset \mathcal{F}_t$, then there exists a process $\{\bar{\mu}_t\}$ such that*

$$\bar{\alpha} := (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{X_t\}, \{\bar{\mu}_t\}, S)$$

is in $\tilde{\mathcal{U}}_{sx}$ and

$$J(s, \bar{\alpha}) = J(s, \alpha). \tag{3.7}$$

Proof. Let $\bar{\mu}_t$ be a progressive version of $P(\mu_t \mid \mathcal{G}_t)$. Such $\bar{\mu}_t$ exists because if $\{\phi^n\}$ is a countable dense subset of $C_b(U)$, then there exists a version $\bar{\mu}_t$ of $P(\mu_t \mid \mathcal{G}_t)$ such that $\langle \bar{\mu}_t, \psi\rangle$ is progressive where ψ is an element of the countable set

$$C_0 := \bigcup_{i=1}^{\infty} A_i \cup \{1\}$$

and A_i consists of all i-fold products of elements of $\{\phi^n\}_{n=1}^{\infty}$. Let H be the vector space of all elements ϕ of $C_b(U)$ such that $\langle \bar{\mu}_t, \phi\rangle$ is progressive. It contains the constant functions and is closed under uniform limits as well as under monotone pointwise limits. By the monotone class theorem [6, Thm. 21, Chap. I] $H = C_b(U)$, so $\bar{\mu}_t$ is progressive.

Observe that

$$\begin{aligned} P\int_s^{T\wedge y} |\bar{\mu}_t|^p\, dt &= P\int_s^T \mathbb{1}_{\{t<y\}} \int_U |u|^p \bar{\mu}_t(du)\, dt \\ &= P\int_s^T \mathbb{1}_{\{t<y\}} P\left\{\int_U |u|^p \mu_t(du) \mid \mathcal{G}_t\right\} dt \\ &= P\int_s^{T\wedge y} |\mu_t|^p\, dt. \end{aligned}$$

Since $\mathcal{F}_t^{X,S} \subset \mathcal{G}_t$ then X_t^* is $\mathcal{G}_t$ measurable and for ϕ in $C_b^2(\mathbb{R}^d)$

$$\begin{aligned} &P\left\{\phi(X_{t+h}^*) - \int_t^{t+h} \mathbb{1}_{\{\theta<y\}} L\phi(\theta, X_\theta, \bar{\mu}_\theta)\, d\theta \mid \mathcal{G}_t\right\} \\ &\quad = P\left\{\phi(X_{t+h}^*) - \int_t^{t+h} \mathbb{1}_{\{\theta<y\}} \int_U L\phi(\theta, X_\theta, u) P(\mu_\theta(du) \mid \mathcal{G}_\theta)\, d\theta \mid \mathcal{G}_t\right\} \\ &\quad = P\left\{\phi(X_{t+h}^*) - \int_t^{t+h} \mathbb{1}_{\{\theta<y\}} L\phi(\theta, X_\theta, \mu_\theta)\, d\theta \mid \mathcal{G}_t\right\} \\ &\quad = P\left\{P\left[\phi(X_{t+h}^*) - \int_t^{t+h} \mathbb{1}_{\{\theta<y\}} L\phi(\theta, X_\theta, \mu_\theta)\, d\theta \mid \mathcal{F}_t\right] \,\middle|\, \mathcal{G}_t\right\} \\ &\quad = P\{\phi(X_t^*) \mid \mathcal{G}_t\} \\ &\quad = \phi(X_t^*). \end{aligned}$$

It follows that $\bar{\alpha}$ is in $\tilde{\mathcal{U}}_{sx}$.

The same kind of conditioning argument can be used to establish (3.7). □

The first choice of a canonic Ω would be to take $\Omega = C = C(\mathbb{R}^+; \mathbb{R}^d)$, the space of X-trajectories. This leads to the natural controls.

DEFINITION 3.8. α in $\mathcal{U}_{sx}$ is a *natural control*, i.e., $\alpha \in \mathcal{N}_{sx}$, if $\Omega = C$, $\mathcal{F} = \mathcal{C}$, $\mathcal{F}_t = \mathcal{C}_t$, $X_t(\omega) = \omega(t)$,

$$P\{\omega(t) = x,\ 0 \leqq t \leqq s\} = 1,$$

and S is a wide sense stopping time. We observe that a natural control is specified by the triple $(P, \{u_t\}, S)$: a probability measure on C (the distribution of $\{X_t\}$), a natural control process (i.e., a progressive function: $\mathbb{R}_+ \times C \to U$), and a random variable $S \geqq s$ that is a stopping time relative to $\{\mathcal{C}_{t+}\}$.

COROLLARY 3.9. *Assume* (3.4) *and* (3.5). *If* $\tilde{\alpha}$ *is in* $\tilde{\mathcal{U}}^f_{sx}$ *such that* $\tilde{S}$ *is a* $\{\mathcal{F}^X_t\}$ *stopping time, then there exists a natural control* α *such that*

$$J^1_i(s, \tilde{\alpha}) \geqq J^1_i(s, \alpha), \qquad J^2_i(s, \tilde{\alpha}) = J^2_i(s, \alpha).$$

Proof. With $\mathcal{G}_t = \mathcal{F}^X_t$ we apply Lemma 3.7 and Theorem 3.6 to

$$\tilde{\alpha} = (\tilde{\Omega}, \tilde{F}, \tilde{P}, \{\mathcal{F}_t\}, \{X_t\}, \{\mu_t\}, \tilde{S})$$

to obtain

$$\bar{\alpha} = (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}^X_t\}, \{X_t\}, \{\bar{u}_t\}, \tilde{S})$$

in $\mathcal{U}^f_{sx}$. Let P be the law of X on C. A standard result implies that there exists a measurable, $\{\mathcal{C}_{t+}\}$ adapted function ϕ such that

$$\bar{u}_t(\omega) = \phi(t, X(\omega)) \quad \text{a.e. a.s.}$$

Now Lemma A.1 ensures the existence of a natural control process $u(t, x)$ "equivalent" to ϕ, hence $\bar{u}$. Similarly, if $Y_t = \mathbb{1}_{\{\tilde{S} \leqq t\}}$, then there exists a $\{\mathcal{C}_{t+}\}$ adapted process ψ_t on $(C, \mathcal{C}, P)$ and a null set N in $\mathcal{F}$ such that for $\omega \notin N$, $Y_t(\omega) = \psi_t(X(\omega))$ almost everywhere (a.e.). If $\Psi_t(x) = \lim_{\varepsilon \to 0+} \operatorname{ess\,sup}_{t < r < t+\varepsilon} \psi_r(x)$, then for $\omega \notin N$, $\Psi_t(X(\omega)) = Y_t(\omega)$ for all t and Ψ is indistinguishable from a $\{\mathcal{C}_{t+}\}$ progressive process $\tilde{\Psi}$ (c.f. [6, IV 37, IV 38]. If $S(x) = \inf\{t \geqq s\colon \tilde{\Psi}_t(x) = 1\}$, then S is a $\{\mathcal{C}_{t+}\}$ stopping time and $S \circ X = \tilde{S}$ almost surely. The result follows. □

We note that if $g = 0$ and $h = 0$, then S only figures in the problem as an upper limit of integration and we can move it into the integrand as, for example,

$$\int_s^{\rho \wedge S} f_i(t, X_t, u_t)\, dt = \int_s^{\rho} \mathbb{1}_{\{t < S\}} f_i(t, X_t, u_t)\, dt.$$

Hence in the definition of a control we can replace S by a progressive decreasing process whose values lie in $\{0, 1\}$ almost everywhere, almost surely. In this way we avoid going to wide sense stopping times for natural controls.

Let $\mathcal{N}^f_{sx}$ denote the set of feasible natural controls with initial condition (s, x). Then $\mathcal{N}^f_{sx} \subset \mathcal{U}^f_{sx} \subset \tilde{\mathcal{U}}^f_{sx}$. If $S = T$ fixed, $s \leqq T \leqq +\infty$ (e.g., $h_0(t, x) = +\infty$ for $t \neq T$) then the above result implies that subject to (3.4), (3.5)

$$\begin{aligned}\inf\{J^1_0(s, \alpha)\colon \alpha \in \mathcal{N}^f_{sx}\} &= \inf\{J^1_0(s, \alpha)\colon \alpha \in \mathcal{U}^f_{sx}\} \\ &= \inf\{J^1_0(s, \alpha)\colon \alpha \in \tilde{\mathcal{U}}^f_{sx}\}.\end{aligned}$$

An approach along these lines was taken by Haussmann [11]—for the natural controls the process $\{u_t\}$ was replaced by a progressive selection of the multifunction $(t, \omega) \to K(t, \omega(t))$. As the method is analogous to the one used in the deterministic theory we expect and find the same technical difficulties in verifying the closure

property. Many of these difficulties can be circumvented by a slightly different method introduced by El Karoui, Huu Nguyen, and Jeanblanc-Picqué [8]. It is the one that we will use here.

Rather than work with natural controls we choose as the canonic Ω the space of trajectories of $(\{X_t\}, \{\mu_t\}, \{\mathbb{1}_{\{S \leqq t\}}\})$.

3.10. The space V. Let V be the set of measurable functions $\eta : \mathbb{R}_+ \to M_1(U)$. We will be working with $M_1(V)$ so we want to put a metric topology on V, and we want it to be such that for all $0 \leqq s < t < \infty$ and for all ϕ in $C_b(U)$

$$\int_s^t \langle \eta_\theta^n, \phi \rangle \, d\theta \to \int_s^t \langle \eta_\theta, \phi \rangle \, d\theta$$

if $\eta^n \to \eta$. Fortunately, we can achieve this, and we can do it in either of two equivalent ways. Let us define $\tilde{\eta}_t$ in $M_+(U)$ by

$$\tilde{\eta}_t(\cdot) = \int_0^t \eta_\theta(\cdot) \, d\theta. \tag{3.8}$$

Then $\tilde{\eta}$ is in $C(\mathbb{R}_+; M_+(U))$, a Polish space. If we set $i(\eta) = \tilde{\eta}$ then $i(V)$ is a closed subset of $C(\mathbb{R}_+; M_+(U))$ by a result of Sion ([18, Thm. 5.1, III]) because the weak topology on $M_1(U)$ is a weak * topology (cf. [19, Thm. 1.1.2]). Hence V under the topology induced by i is a Polish space. Since $i(V)$ is an equicontinuous family, then it follows from Ascoli's Theorem that a set A in V is sequentially compact if for each $t < \infty$

$$\left\{ \int_0^t \eta_\theta \, d\theta : \eta \in A \right\} \subset M_+(U)$$

is sequentially compact, i.e., if for any $\delta > 0$ there exists a compact set K in U such that for all $\eta \in A$

$$\int_0^t \eta_\theta(K) \, d\theta > t - \delta.$$

Alternatively if η is in V, then the set function $\bar{\eta}$ defined on $\mathcal{R}_+ \times \mathcal{U}$ by

$$\bar{\eta}(A \times B) = \int_A \eta_t(B) \, dt \tag{3.8'}$$

is an element of $M_+(\mathbb{R}_+ \times U)$. The stable topology on $M_+(\mathbb{R}_+ \times U)$ is the weakest topology, which renders continuous the mappings

$$\bar{\eta} \to \int_{\mathbb{R}_+} \int_U \phi(t, u) \bar{\eta}(dt, du)$$

for all bounded, real-valued, measurable ϕ that are continuous in u and that satisfy $\phi(t, u) = 0$ if $t > T_\phi$ for some constant T_ϕ. Jacod and Mémin [13] have shown that $M_+(\mathbb{R}_+ \times U)$ with the stable topology is a separable metric space and they have characterized the sequentially compact sets. If $\bar{i}(\eta) = \bar{\eta}$, then $\bar{i}(V)$ is a closed subset of $M_+(\mathbb{R}_+ \times U)$ (stable topology) because $\mathbb{R}_+ \times U$ is Polish so that we can always disintegrate. If we now put the topology induced by $\bar{i}$ on V, then it is clear that we obtain exactly the same topology as introduced above (cf. (3.8) and (3.8')). The criterion for sequential compactness given above can now also be derived from Theorem 2.8 of [13].

The canonic filtration on V is $\{\mathcal{V}_t\}$ where $\mathcal{V}_t$ is generated by sets of the form

$$\left\{\eta: \int_0^s \eta_\theta \, d\theta \in B\right\}$$

with $s \leqq t$ and B a Borel set in $M_1(U)$.

Recall that we had extended u_t to $\mathbb{R}_+$ by setting $u_t = u^0$ for $t < s$ for some fixed $u^0 \in U$. If δ^0 is the point mass at u^0, i.e., is the Dirac measure at u^0, then μ_t can be defined for $0 \leqq t < s$ as $\mu_t = \delta^0$.

3.11. The space Z. S assumes its values in $\bar{\mathbb{R}}_+$, which is a compact metric space with metric

$$d(x, x') = |r(x) - r(x')|, \qquad r(x) = x/(1+x).$$

We can identify $\bar{\mathbb{R}}_+$ with a space of functions by noting that if $\Delta \in \bar{\mathbb{R}}_+$ and

$$\zeta(t) = \mathbb{1}_{\{t \geqq \Delta\}},$$

then ζ lies in Z, the set of all distribution functions of Dirac point measures on $\bar{\mathbb{R}}_+$. We denote the map $\zeta \to \Delta$ by $\Delta(\cdot)$. It maps Z into $\bar{\mathbb{R}}_+$. The topology inherited by Z from $\bar{\mathbb{R}}_+$ via the map $\Delta(\cdot)$ is that of convergence at all points of continuity and at ∞. We could also add that it is the topology of weak convergence of the corresponding (point) probability measures. In any case, Z is a compact metric space with canonic filtration

$$\mathcal{Z}_t = \sigma[\zeta(\theta);\ \theta \leqq t].$$

In $\bar{\mathbb{R}}_+$ the corresponding σ-algebra is

$$(\bar{\mathcal{R}}_+)_t = \sigma[[0, \theta];\ \theta \leqq t],$$

i.e., sets of the form B or $B \cup (t, \infty]$ where B is a Borel subset of $[0, t]$.

We say that α in $\tilde{\mathcal{U}}_{sx}$ is a *canonic relaxed control* if

$$\Omega = C \times V \times Z, \quad \mathcal{F} = \mathcal{Q}, \quad \mathcal{F}_t = \mathcal{C}_t \times \mathcal{V}_t \times \mathcal{Z}_t,$$

$$X_t(\omega) = \xi(t), \quad \mu_t(\omega) = \eta(t), \quad S(\omega) = \Delta(\zeta)$$

with $\omega = (\xi, \eta, \zeta)$ and if

$$P\{\omega(t) = (x, \delta^0, 0),\ 0 \leqq t < s\} = 1.$$

It follows that a canonic relaxed control is completely specified by the probability P on $C \times V \times Z$, i.e., by the distribution of $(\{X_t\}, \{\mu_t\}, \{\mathbb{1}_{\{S \leqq t\}}\})$. We formalize this by making an equivalent definition first proposed by El Karoui, Huu Nguyen, and Jeanblanc-Picqué [8].

DEFINITION 3.12. Given the initial condition (s, x) in D, P is a *control rule*, i.e., P is in $R(s, x)$, if P is a probability measure on $(C \times V \times Z, \mathcal{C} \times \mathcal{V} \times \mathcal{Z})$, such that

$$P\left(\int_0^{T \wedge y} \int_U |u|^p \eta(t; du)\, dt\right) < \infty \quad \text{for all } T < \infty, \tag{3.9}$$

$$M_t^* \phi := M_{t \wedge y} \phi \text{ is a } (P, \mathcal{C}_t \times \mathcal{V}_t \times \mathcal{Z}_t) \text{ martingale on } t \geqq s \text{ for } \phi \text{ in } C_b^2(\mathbb{R}^d) \tag{3.10}$$

where

$$M_t \phi = \phi(\xi(t)) - \int_s^t L\phi(\theta, \xi(\theta), \eta(\theta))\, d\theta,$$

$$P\{\omega(t) = (x, \delta^0, 0);\ 0 \leqq t < s\} = 1. \tag{3.11}$$

Here $y = \rho(\xi) \wedge \Delta(\zeta)$, $\rho(\xi)$ is the first exit time after s of $(t, \xi(t))$ from D and $\Delta(\zeta)$ is the time when $\zeta(t)$ jumps from zero to one.

Observe now that given P in $R(s, x)$ and setting $\Omega = C \times V \times Z$, $\mathcal{A}_t = \mathcal{C}_t \times \mathcal{V}_t \times \mathcal{Z}_t$, we have a canonic relaxed control $(\Omega, \mathcal{A}, P, \{\mathcal{A}_t\}, \{\xi(t)\}, \{\eta(t)\}, \Delta(\zeta))$ in $\tilde{\mathcal{U}}_{sx}$. We define

$$F_i(s, \omega) = \int_s^y f_i(\theta, \xi(\theta), \eta(\theta))\, d\theta + h_i(y, \xi(y)),$$

$$G_i(\omega) = g_i(y, \xi(y)),$$

and for $P \in R(s, x)$, $J_i^1(s, P) = PF_i(s, \omega)$ and, similarly, $J_i^2(s, P) = PG_i(\omega)$. Now $R^f(s, x)$, the set of *feasible rules*, is defined in the obvious way.

The final result of this section shows that in the control problem we may restrict attention to control rules. Clearly, $R^f(s, x) \hookrightarrow \tilde{\mathcal{U}}_{sx}^f$.

THEOREM 3.13. *If α is in $\tilde{\mathcal{U}}_{sx}^f$, then there exists $\bar{P}$ in $R^f(s, x)$ such that $J(s, \alpha) = J(s, \bar{P})$.*

Proof. In Lemma 3.7 take $\mathcal{G}_t = \mathcal{F}_t^{X,\mu,S}$. Then

$$\bar{\mu}_t = P(\mu_t \mid \mathcal{G}_t) = \mu_t \quad \text{a.s.}$$

so $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t^{X,\mu,S}\}, \{X_t\}, \{\mu_t\}, S)$ is in $\tilde{\mathcal{U}}_{sx}^f$. The result follows if we take the image under

$$(X, \mu, \mathbb{1}_{\{S \leqq \cdot\}}) : \Omega \to C \times V \times Z. \qquad \square$$

Remark 3.14. We point out that a canonic relaxed control is not a natural control, that is, there may still be some randomization that is not X-measurable since in most cases $\mathcal{F}_t^{X,\mu,S} \neq \mathcal{F}_t^X$. However we can say that

$$\inf\{J_0^1(s, \alpha) : \alpha \in \tilde{\mathcal{U}}_{sx}^f\} = \inf\{J_0^1(s, P) : P \in R^f(s, x)\}.$$

Moreover, if the infimum is attained in any one of $\tilde{\mathcal{U}}_{sx}^f$, $R^f(s, x)$, $\mathcal{U}_{sx}^f$, then it is attained in the other two also. If S is an $\{\mathcal{F}_t^X\}$ stopping time (e.g., $h_0(t, x) = +\infty$ for $t \neq T$), then we can add $\mathcal{N}_{sx}^f$ to the above list.

4. Existence of optimal controls. We will now show that if

$$R'(s, x) = \operatorname{argmin}\{J_0^1(s, P) : P \in R^f(s, x)\},$$

then $R'(s, x)$ is not empty provided certain hypotheses are met (notably (3.5)). According to the previous section, this implies that there exists an optimal relaxed control, and if (3.4) also holds, then there exists an optimal (strict) control. The argument is straightforward: a lower semicontinuous (l.s.c.) function attains its infimum on a nonempty compact set. The first step is to show that $P \to J_i^1(s, P)$ is lower semicontinuous and $P \to J_i^2(s, P)$ is continuous at least under some reasonable hypotheses. We write $\bar{D}_t$ for the t-section of $\bar{D}$. Let $E \subset R(s, x)$ be such that $J_i^1(s, P) \leqq \lambda$ for all $P \in E$, some $\lambda < \infty$ (l is given in (3.5)).

LEMMA 4.1. *Assume that*

(i) *$\xi \to \rho(\xi)$ is continuous as a mapping into $\bar{\mathbb{R}}_+$, P-a.s. for all $P \in E$,*

(ii) $\lim_{t \to \infty} \sup_{x \in \bar{D}_t} |g(t, x)| < \infty$,

(iii) $|g(t, x)| \leqq k(1 + |x|^\alpha)$, $\quad 0 \leqq \alpha < \bar{p}$.

Then $P \to J_i^2(s, P)$ is continuous on E, $i = 1, 2, \cdots, n$.

Proof. The continuity of g and (i) imply that G is almost surely continuous. Moreover, for t sufficiently large, i.e., $t > T$, $g(t, x)$ is bounded (cf. (ii)), hence by (iii) and Lemma 3.3,

$$P\{|G|^{\bar{p}/\alpha}\} \leqq \bar{k}_1(1 + P\{\|\xi\|_T^{\bar{p}}\}) < \bar{k}_2$$

where $\bar{k}_2$ depends on T, $\xi(s) = x$ and λ but is the same for all P in E.

It follows that

$$\lim_{N\to\infty} \sup_{P\in E} P\{|G|\mathbb{1}_{\{|G|>N\}}\} = 0,$$

and hence for all i

$$\lim_{N\to\infty} \sup_{P\in E} |P(G_i \wedge N) - P(G_i)| = 0.$$

Now let $P_n \to P$. The continuity of G implies that

$$\lim_n P_n(G_i \wedge N) = P(G_i \wedge N)$$

so by the above and monotone convergence

$$\lim_n P_n(G_i) = \lim_N \lim_n P_n(G_i \wedge N) = \lim_N P(G_i \wedge N) = P(G_i).$$

Thus $J_i^2(s, \cdot)$ is continuous. □

Remark 4.2. Let us discuss briefly the continuity of ρ. If $D = [0, T) \times \mathbb{R}^d$ with $0 < T \leqq \infty$ then $\rho = T$, and hence it is continuous. More generally if $D = [0, T) \times O$ with $0 < T \leqq \infty$ and O has smooth boundary ∂O, we assume that ∂O is the union of three sets Γ_0, Γ_1, Γ_2. With $\delta(x) := \operatorname{dist}(x, \partial O)$ we assume that Γ_0 has a neighbourhood N_0 such that on $[0, T) \times (N_0 \cap O) \times U$

$$\alpha(t, x, u) := \sum_{ij} a_{ij}(t, x, u)\partial_i \delta(x)\partial_j \delta(x) = 0,$$

$$\beta(t, x, u) := \sum_i b_i(t, x, u)\partial_i \delta(x) + \frac{1}{2}\sum_{ij} a_{ij}(t, x, u)\partial_i \partial_j \delta(x) \geqq 0,$$

where $\partial_i := \partial/\partial x_i$. Then $P\{\xi(\rho) \in \Gamma_0\} = 0$, i.e., $(t, \xi(t))$ does not exit D through $[0, T) \times \Gamma_0$, P-a.s. for any rule P. In fact, if $\xi(t) \in N_0$ then

$$\frac{d\delta(\xi_t)}{dt} = \beta(t, \xi_t, \eta_t) \geqq 0 \quad P\text{-a.s.}$$

so that δ cannot decrease to zero, i.e., $\xi(t)$ cannot approach Γ_0 in finite time.

We also assume that for some $\varepsilon > 0$ Γ_1 has an ε-neighbourhood N_1 such that on $[0, T) \times (N_1 \cap O) \times U$

$$\alpha(t, x, u) = 0, \qquad \beta(t, x, u) \leqq -\nu$$

for some $\nu > 0$. Now

$$\frac{d\delta(\xi_t)}{dt} \leqq -\nu,$$

i.e., if $t' \geqq t$, then

$$t' - t \leqq \nu^{-1}[\delta(\xi_t) - \delta(\xi_{t'})].$$

If $\xi^n \to \xi$ and $\rho^n := \rho(\xi^n) < \rho := \rho(\xi)$, then $\rho^n \to \rho$ since $\rho(\cdot)$ is l.s.c. If $\rho^n > \rho$, then for n sufficiently large $\xi^n(\rho) \in N_1$, so $\delta(\xi_t)$ decreases as t increases, i.e., $\xi^n(t) \in N_1$ for $t \geqq \rho$. Moreover,

$$\rho^n - \rho \leqq \nu^{-1}\delta[\xi^n(\rho)] \leqq \nu^{-1}|\xi^n(\rho) - \xi(\rho)| \to 0.$$

Hence ρ is continuous at ξ such that $\xi(\rho) \in \Gamma_1$.

Finally, we assume that Γ_2 has neighbourhood N_2 such that on $[0, T) \times (N_2 \cap O) \times U$,

$$\alpha(t, x, u) \geqq \nu > 0.$$

Using a Girsanov transformation and a time change, we can map $\delta[\xi(t)]$ into a Brownian motion $\tilde{W}$ under a measure Q equivalent to P. Now ρ is the continuous image of the first exit time of $\tilde{W}$ from $\mathbb{R}_+$, and hence is Q-a.s. (and thus P-a.s.) continuous at ξ if $\xi(\rho) \in \Gamma_2$. Since all cases are now covered, ρ is continuous.

We obtain a somewhat better result using more complicated arguments in Theorem A12 of the Appendix.

Remark 4.3. There are two cases where it is readily seen that $\omega \to h_i(y, \xi(y))$ is l.s.c. P-a.s. If ρ is continuous P-a.s., then $h_i(y, \xi(y))$ is l.s.c. since h_i is. Alternatively, if $D = [0, T) \times O$ with O open, then ρ is l.s.c. If $h_i(t, x) = \tilde{h}(t)$ for x in the boundary of O with $\tilde{h}$ left-continuous and nondecreasing, then $\tilde{h}(\rho)$ is l.s.c. Since also h_i is l.s.c. and y is continuous on $D \cup \{T\} \times O$, then $h_i(y, \xi(y))$ is l.s.c.

LEMMA 4.4. *Assume that $\omega \to h_i(y, \xi(y))$ is* l.s.c. *P-a.s. for all $P \in E$ and all i. Then $P \to J_i^1(s, P)$ is* l.s.c. *on E for all i.*

Proof. Let us first show that F_i is l.s.c. Assume $(\xi^n, \eta^n, \zeta^n) \to (\xi^0, \eta^0, \zeta^0)$. Note that the function $y(\xi, \zeta) = \rho(\xi) \wedge \Delta(\zeta)$ is l.s.c. since ρ and Δ are. Define $y^n = y(\xi^n, \zeta^n)$. For τ in $(0, \infty)$ define

$$z = \begin{cases} y^0 & \text{if } y^0 < \infty, \\ \tau & \text{otherwise.} \end{cases}$$

Fix $N < \infty$ and write

$$f_N^n(t, u) = f_i(t, \xi_t^n, u) \wedge N,$$

$$\phi^n(t) = \begin{cases} 1 & \text{if } y^n \leqq t \leqq y^0, \\ 0 & \text{otherwise,} \end{cases}$$

$$dQ^n = \eta_t^n(du)\, dt.$$

For $y^0 < \infty$

$$\int_{y^0}^{y^n} f_i(t, \xi_t^n, \eta_t^n) \wedge N\, dt \geqq -\int \phi^n(t) f_N^n(t, u)\, dQ^n$$

and for any $\bar{\gamma} > 0$

$$Q^n\{(t, u): |\phi^n(t) f_N^n(t, u)| > \bar{\gamma}\} \leqq (y^0 - y^n)^+.$$

Since y is l.s.c., $\liminf_n y^n \geqq y^0$, so

$$\lim_n (y^0 - y^n)^+ = 0.$$

By Lemma A.2(iii),

$$\lim_n \int \phi^n(t) f_N^n(t, u)\, dQ^n = 0,$$

so

$$\liminf_{n \to \infty} \int_{y^0}^{y^n} f_i(t, \xi_t^n, \eta_t^n)\, dt \geqq 0.$$

If $y^0 = +\infty$ then $y^n \to \infty$ since y is l.s.c. so

$$\liminf_{n \to \infty} \int_{\tau}^{y^n} f_i(t, \xi_t^n, \eta_t^n)\, dt \geqq 0.$$

Hence we always have

$$\begin{aligned}
\liminf_{n\to\infty} \int_s^{y^n} f_i(t, \xi_t^n, \eta_t^n)\, dt &\geqq \liminf_{n\to\infty} \int_s^z f_i(t, \xi_t^n, \eta_t^n)\, dt \\
&\geqq \liminf_{n\to\infty} \int_s^z f_i(t, \xi_t^n, \eta_t^n) \wedge N\, dt \\
&= \liminf_{n\to\infty} \int_s^z \int_U f_N^n\, dQ^n \qquad (4.1) \\
&\geqq -\limsup_{n\to\infty} \int_s^z \int_U (f_N^n - f_N^0)^-\, dQ^n + \liminf_{n\to\infty} \int_s^z \int_U f_N^0\, dQ^n \\
&\geqq -\limsup_{n\to\infty} \int_s^z \int_U (f_N^n - f_N^0)^-\, dQ^n + \int_s^z \int_U f_N^0\, dQ
\end{aligned}$$

where the last inequality follows because $Q^n \to Q$ in the stable topology and $f_N^0(t, \cdot)$ is l.s.c. Note that Dellacherie and Meyer [6, Thm. 55, Chap. III] show that

$$\liminf_{n\to\infty} Q_n\phi \geqq Q\phi$$

if $Q_n \to Q$ is the topology of weak convergence and if $\phi \geqq 0$, l.s.c., whence the same result follows for stable convergence (if $\phi(t, \cdot)$ is l.s.c.) by Fatou's Lemma.

From Lemma A.3(i) with $\bar{\phi}(t, x, u) = f_i(t, x, u) \wedge N$, we conclude that

$$\lim_n \int_s^z \int_U (f_N^n - f_N^0)^- dQ^n = 0,$$

so from (4.1) and monotone convergence we conclude that

$$\liminf_{n\to\infty} \int_s^{y^n} f_i(t, \xi_t^n, \eta_t^n)\, dt \geqq \int_s^z f_i(t, \xi_t, \eta_t)\, dt. \qquad (4.2)$$

Recalling that h_i is l.s.c. then (4.2) implies the same for F_i in case $y^0 < +\infty$. If $y^0 = +\infty$ then $z = \tau$ and we simply take the limit as $\tau \to \infty$ in (4.2).

Hence F_i is l.s.c. P-a.s., $F_i \geqq 0$. Then there exists a P-null set N and bounded functions ϕ_m such that ϕ_m is continuous on $\Omega \backslash N$ and $\phi_m \uparrow F_i$ on $\Omega \backslash N$, $\phi_m \geqq 0$. Now if $P_n \to P$ then by monotone convergence

$$P(F_i) = \lim_m P(\phi_m) = \lim_m \lim_n P_n(\phi_m) \leqq \lim_m \liminf_n P_n(F_i) = \liminf_n P_n(F_i)$$

so that J_i^1 is l.s.c. □

Let us now find a suitable compact set on which J_0^1 can attain its inf. Unfortunately, $R^f(s, x)$ is not compact—the difficulty lies in the fact that in (3.9) we have $T \wedge y$ rather than T and in (3.10) we have $M_t^*\phi$ rather than $M_t\phi$. The following technical device will allow us to overcome this problem. Let us adjoin $\{u^*\}$ to U where u^* is an isolated point, for example, if U is in $\mathbb{R}^l$ then $U \cup \{u^*\}$ is isomorphic to $\{(u, 0): u \in U\} \cup \{(0, 1)\}$ in $\mathbb{R}^{l+1}$. We define

$$a(t, x, u^*) = 0, \quad b(t, x, u^*) = 0, \quad f_i(t, x, u^*) = +\infty, \quad \tilde{f}(u^*) = 0.$$

From now on we assume that such u^* is in U. This does not affect $K(t, x)$ since this set contains only u such that $f_i < \infty$. Let δ^* be the point mass at u^*. Observe that for any (feasible) control in $\tilde{\mathcal{U}}_{sx}^f$, u^* is not in the support of μ_t for $t \leqq y$ except possibly on a null set. On the other hand, if we have a (feasible) control α in $\tilde{\mathcal{U}}_{sx}^f$ and if we redefine $\mu_t = \delta^*$, $t > y$, then we have replaced X_t by $X_t^* = X_{t\wedge y}$ and α by α^*, but $J(\alpha^*) = J(\alpha)$.

In terms of control rules, given P in $R(s, x)$ we replace it by P_0 in $R(s, x)$ where $P_0 = P$ on $\mathcal{A}_y$, i.e., on the events that happen prior to time y, and

$$P_0 = \delta_{\xi(y(\omega))} \times \delta^* \times \delta_{y(\omega)}$$

on $\mathcal{A}^y$, the events that happen after time $y(\omega)$. Here $\delta_{\xi(y(\omega))}$ is the Dirac measure at the constant function $x(t) = \xi(y(\omega))$ in $C(\mathbb{R}_+; \mathbb{R}^d)$ and $\delta_{y(\omega)}$ is the Dirac measure at $\mathbb{1}_{\{t \geqq y(\omega)\}}$ in Z. The techniques of Stroock and Varadhan [19, § 6.1] (cf. § 5.6 of the present paper) allow us to conclude that P_0 is well defined. Now P_0 satisfies the following:

$$P_0 \int_0^T |\eta(t)|^p \, dt < \infty \quad \forall T < \infty \tag{3.9'}$$

$$M_t\phi \text{ is a } (P_0, \underline{\Omega}_t) \text{ martingale for } t \geqq s, \tag{3.10'}$$

$$\begin{aligned} &P_0\{\omega\colon \omega(t) = (x, \delta^0, 0), 0 \leqq t \leqq s\} = 1, \\ &P_0\{\omega\colon (\xi(t), \eta(t)) = (\xi(y(\omega)), \delta^*), t > y(\omega)\} = 1. \end{aligned} \tag{3.11'}$$

Since $F_i(s, \omega)$ and $G_i(\omega)$ are $\mathcal{A}_y$ measurable, $J(P) = J(P_0)$. Let us define $i_0 P = P_0$ and

$$\begin{aligned} R_0(s, x) &= i_0 R(s, x) \\ &= \{P \in M_1(C \times V \times Z)\colon (3.9'), (3.10'), (3.11') \text{ hold}\} \end{aligned}$$

and let $R_0^f(s, x) = i_0 R^f(s, x)$. We conclude that

$$\inf\{J_0^1(s, P)\colon P \in R^f(s, x)\} = \inf\{J_0^1(s, P)\colon P \in R_0^f(s, x)\}. \tag{4.3}$$

We will then find a subset E of $R_0^f(s, x)$ such that

$$\inf\{J_0^1(s, P)\colon P \in R_0^f(s, x)\} = \inf\{J_0^1(s, P)\colon P \in E\}$$

and such that its closure $\bar{E}$ is compact and contained in $R^f(s, x)$. It follows that if $J_0^1(s, \cdot)$ is l.s.c. on $\bar{E}$, then all the above infima are equal and are attained in $\bar{E}$, and hence in $R^f(s, x)$ and even in $R_0^f(s, x)$.

In the case where U is compact, the above construction is unnecessary (note U compact implies that we may take $p = \gamma = \nu = 0$). We can extend a, b to $\mathbb{R}_+ \times \mathbb{R}^d \times U$ so that (2.1) still holds. In (C_3') let us now assume that $M_t(\phi, \alpha)$ is a $(P, \{\mathcal{F}_t\})$ martingale for $t \geqq s$. Note that now (2.2) holds for all $t \geqq s$ and (2.3) holds with X^* replaced by X. The rules corresponding to this modification are:

$$R_c(s, x) = \{P \in M_1(C \times V \times Z)\colon (3.10'), (3.11) \text{ hold}\}.$$

Of course since $p = 0$ then (3.9) and (3.9') are trivial and

$$R_0(s, x) \subset R_c(s, x) \subset R(s, x).$$

Let us define

$$R_0'(s, x) = \operatorname{argmin}\{J_0^1(s, P)\colon P \in R_0^f(s, x)\} = i_0 R'(s, x).$$

With l as in (3.5) let E be a subset of $R_0(s, x)$ such that for some $\lambda < \infty$

$$\sup\{J_l^1(s, P)\colon P \in E\} \leqq \lambda.$$

PROPOSITION 4.5. *Assume* (3.5). *Then E is tight.*

Proof. The result will follow from Theorem A.8 once we have established (A2) and (A4). From (3.5) it follows that for any $\bar{\gamma} > 0$ there exists m such that $\nu_m \leqq \bar{\gamma}$. Now

$$|\eta_t|^p \leqq m^p + \bar{\gamma}\tilde{f}(\eta_t)$$

for all t. We set $q_t(\eta)=\tilde{f}(\eta_t)$. Then from (3.5), (3.11'), and Lemma 3.3 for any $T<\infty$ and sufficiently large M (cf. the Appendix for the definition of Σ_M^c)

$$\begin{aligned}
P\{\Sigma_M^c\} &\leqq P\{\|\xi\|_T \geqq M\} + P\left\{\int_0^T |\eta_\theta|^p \, d\theta \geqq M\right\} + P\left\{\int_0^T \tilde{f}(\eta_\theta)\, d\theta \geqq M\right\} \\
&\leqq P\{\|\xi\|_T \geqq M\} + P\left\{\int_s^y |\eta_\theta|^p \, d\theta + T\kappa_1^p \geqq M\right\} + P\left\{\int_s^y \tilde{f}(\eta_\theta)\, d\theta + T\kappa_2 \geqq M\right\} \\
&\leqq \frac{K(T)}{M}\,(J_l^1(s,P)+|x|+1) \\
&\leqq K(T)(\lambda+|x|+1)/M
\end{aligned}$$

for some constant $K(T)$. Here $\kappa_1=\max\{|u^0|,|u^*|\}$ and $\kappa_2=\max\{\tilde{f}(u^0),\tilde{f}(u^*)\}=\tilde{f}(u^0)$. The last expression tends to zero uniformly in $P\in E$ as $M\to\infty$. Hence (A2) is established.

In the case where $p=0$ we take $\bar{\gamma}=0$ in the above and dispense wth q_t. Then

$$\begin{aligned}
P\{\Sigma_M^c\} &\leqq P\{\|\xi\|_T \geqq M\} + P\left\{T\int_0^T \eta_\theta(|u|>M)\, d\theta \geqq 1\right\} \\
&\leqq P\{\|\xi\|_T \geqq M\} + P\left\{\int_0^T \int_{\{|u|>M\}} \tilde{f}(u)\eta_\theta(du)\, d\theta \geqq \frac{k_M}{T}\right\} \\
&\leqq P\{\|\xi\|_T \geqq M\} + P\left\{\int_0^T \tilde{f}(\eta_\theta)\, d\theta \geqq \frac{k_M}{T}\right\} \\
&\leqq K(T)\{(1+|x|)/M + J_l^1(s,P)/k_M\}
\end{aligned}$$

where

$$k_M=\inf\{\tilde{f}(u)\colon |u|>M\}\to\infty$$

as $M\to\infty$, so again (A2) holds.

If U is compact and $P\in R_c(s,x)$, then

$$P\left(T\int_0^T \eta_\theta(|u|>M)\, d\theta \geqq 1\right)=0$$

for M sufficiently large, and we do not require a bound involving $J_l^1(s,P)$, i.e., we do not have to fix P on $\mathcal{U}^y$ to obtain compactness. This explains why we can admit R_c rather than R_0 in this case.

To establish (A4) take ϕ in $C_b^2(\mathbb{R}^d)$. For any P in E, $t\geqq s$ $\{\phi(\xi_t)-\int_s^t L\phi(\theta,\xi_\theta,\eta_\theta)\, d\theta\}$ is a $(P,\mathcal{U}_t)$ martingale, hence so is $\{\phi(\xi_t^M)-\int_s^{t\wedge\sigma_M} L\phi(\theta,\xi_\theta,\eta_\theta)\, d\theta\}$ (cf. the Appendix for the definition of ξ^M). It follows that for $t\geqq s$ and suitable A_ϕ defined below (writing σ for σ_M)

$$\begin{aligned}
&P\left\{\phi(\xi_{t+h}^M)-\phi(\xi_t^M)+A_\phi\int_{t\wedge\sigma}^{(t+h)\wedge\sigma}(1+|\eta_\theta|^p)\, d\theta \,|\, \mathcal{U}_t\right\} \\
&\quad = P\left\{\int_{t\wedge\sigma}^{(t+h)\wedge\sigma}[L\phi(\theta,\xi_\theta,\eta_\theta)+A_\phi(1+|\eta_\theta|^p)]\, d\theta \,|\, \mathcal{U}_t\right\} \\
&\quad \geqq 0
\end{aligned}$$

since by (3.2) for $\theta\leqq\sigma$

$$\begin{aligned}
|L\phi(\theta,\xi_\theta,\eta_\theta)| &\leqq |a(\theta,\xi_\theta,\eta_\theta)||\phi_{xx}(\xi_\theta)|+|b(\theta,\xi_\theta,\eta_\theta)||\phi_x(\xi_\theta)| \\
&\leqq (\|\phi_{xx}\|+\|\phi_x\|)k(1+M^{\bar{\beta}}+|\eta_\theta|^p) \\
&\leqq A_\phi(1+|\eta_\theta|^p)
\end{aligned}$$

if $\bar{\beta} = \max\{1, \beta\}$ and $A_\phi = (\|\phi_{xx}\| + \|\phi_x\|)k(1+M^\beta)$. Clearly, if $\psi(x) = \phi(x-a)$ then $A_\psi = A_\phi$. The same conclusion holds for $t<s$ trivially for any $A_\phi \geqq 0$. Hence (A4) holds and the result follows. □

For $\lambda_0^1 \in (0, \infty)$ define

$$E(\lambda_0^1) = \{P \in R_0^f(s,x): J_0^1(s,P) \leqq \lambda_0^1\}$$

so that $J_i^1(s,P) \leqq \lambda_i^1$, $i = 0, 1, \cdots, m$ for $P \in E(\lambda_0^1)$. Note the inclusion of the case $i=0$. By the above result $E(\lambda_0^1)$ is precompact (take $\lambda = \lambda_l^1$ with l as in (3.5)). We wish to show that its closure $\bar{E}(\lambda_0^1)$ lies in $R^f(s,x)$.

PROPOSITION 4.6. *Assume* (3.5) *and that* J_i^1 *is* l.s.c. *on* $\bar{E}(\lambda_0^1)$ $i = 0, 1, \cdots, m$, *and* J_j^2 *is continuous on* $\bar{E}(\lambda_0^1)$ $j = 1, 2, \cdots, n$. *Then* $\bar{E}(\lambda_0^1) \subset \{P \in R^f(s,x): J_0^1(s,P) \leqq \lambda_0^1\}$.

Proof. Let $\{P_n\}$ be a sequence in $E(\lambda_0^1)$ such that $P_n \to P$ weakly. We wish to show that $P \in R^f(s,x)$ so we must verify that P satisfies (3.9), (3.10), (3.11), and

$$J_i^1(s,P) \leqq \lambda_i^1, \qquad J_i^2(s,P) = \lambda_i^2. \tag{4.4}$$

In fact, (4.4) follows trivially from the same properties for P_n and from the continuity properties of J_i^1, J_i^2. Let

$$A = \{\omega: \omega(t) = (x, \delta^0, 0), 0 \leqq t < s\}.$$

It is a closed set, hence $P(A) \geqq \limsup P_n(A) = 1$, so (3.11) holds. Note that we cannot show that (3.11′) holds (unless ρ is continuous); that is why $E(\lambda_0^1)$ is not closed. Next consider (3.9) for $p>0$ since it is trivial if $p=0$. Fix T and N and define

$$\phi_N(\eta) = \int_0^T \int_U (|u| \wedge N)^p \eta_t(du)\, dt.$$

Since $u \to (|u| \wedge N)^p$ is in $C_b(U)$, then $\eta \to \phi_N(\eta)$ is continuous as a function on V. If $m > \kappa_1$ (cf. the proof of Proposition 4.5 for κ_1), then (3.5) implies that

$$\begin{aligned} P_n\{\phi_N(\eta)\} &\leqq P_n\left\{\int_0^T |\eta_t|^p\, dt\right\} \\ &\leqq m^p T + \nu_m \lambda_l^1. \end{aligned} \tag{4.5}$$

Since ϕ_N is in $C_b(U)$, then by passing to the limit as $n \to \infty$ in (4.5) we obtain

$$P\{\phi_N(\eta)\} \leqq m^p T + \nu_m \lambda_l^1$$

and now by monotone convergence as $N \to \infty$

$$P\left\{\int_0^T |\eta_t|^p\, dt\right\} \leqq m^p T + \nu_m \lambda_l^1. \tag{4.6}$$

Hence P satisfies (3.9′) and then also (3.9).

Now we address (3.10). Fix $s \leqq T < \infty$. With ϕ in $C_c^2(\mathbb{R}^d)$ we write

$$\begin{aligned} M_t\phi &= M_t^{1m} + M_t^{2m}, \\ M_t^{1m} &= \phi(\xi_t) - \int_s^t \int_{|u|<m} L\phi(\theta, \xi_\theta, u)\eta_\theta(du)\, d\theta, \\ M_t^{2m} &= -\int_s^t \int_{|u|\geqq m} L\phi(\theta, \xi_\theta, u)\eta_\theta(du)\, d\theta. \end{aligned}$$

Then M_t^{1m} is bounded on Ω (cf. (2.1)), and for a sequence $m_k \to \infty$, it is P-a.s. continuous in ω. This last point follows from the second part of Lemma A.3 with

$$\bar{\phi}(\theta, x, u) = L\phi(\theta, x, u)$$

and with m_k chosen so that $m_k \to \infty$ and

$$P\int_s^T \eta_\theta(|u| = m_k)\, d\theta = 0.$$

It follows that for any bounded continuous function $H: \Omega \to \mathbb{R}$, and any k,

$$\lim_n P_n(M_t^{1m_k} H) = P(M_t^{1m_k} H)$$

and similarly for t replaced by T.

Now consider M_t^{2m}. If $m > \kappa_1$ then there is a constant $\bar{A}_\phi$ such that

$$|M_t^{2m}| \leqq \bar{A}_\phi \left[\int_s^t \eta_\theta(|u| \geqq m)\, d\theta + \nu_m \int_s^{t\wedge\rho} f_l(\theta, \xi_\theta, \eta_\theta)\, d\theta\right] \quad P_n\text{-a.s.} \tag{4.7}$$

with $\nu_m \to 0$ as $m \to \infty$ (cf. (3.5), (3.11′)). Since $\{P_n\}$ is tight (since $E(\lambda_0^1)$ is), given $\gamma > 0$ there exists a compact set K_γ in V such that for all P_n and P, $P_n(K_\gamma) \geqq 1 - \gamma$. Moreover, K_γ compact means that for some m_γ

$$\int_s^T \eta_\theta(|u| \geqq m_\gamma)\, d\theta < \gamma(T - s)$$

for all η in K_γ. Hence

$$\begin{aligned} P_n\left\{\int_s^t \eta_\theta(|u| \geqq m)\, d\theta\right\} &\leqq \gamma(T-s) + P_n\left\{\mathbb{1}_{K_\gamma}\int_s^t \eta_\theta(|u| \geqq m)\, d\theta\right\} \\ &\leqq 2\gamma(T-s) \end{aligned} \tag{4.8}$$

if $m \geqq m_\gamma$, and the same is true with P_n replaced by P and/or t replaced by T.

Hence if H is an element of $C_b(\Omega)$ that is $\mathcal{A}_t$ measurable, then

$$\begin{aligned} |P\{(M_T\phi - M_t\phi)H\}| \leqq\ & |P\{(M_T^{1m} - M_t^{1m})H\} - P_n\{(M_T^{1m} - M_t^{1m})H\}| \\ & + |P\{(M_T^{2m} - M_t^{2m})H\}| + |P_n\{(M_T^{2m} - M_t^{2m})H\}|, \end{aligned} \tag{4.9}$$

since $M\phi$ is a P_n martingale. From (4.7), (4.8) and

$$P_n\left\{\nu_m \int_s^{t\wedge\rho} f_l\, d\theta\right\} \leqq \nu_m \lambda_l^1,$$

$$P\left\{\nu_m \int_s^{t\wedge\rho} f_l\, d\theta\right\} \leqq \nu_m \lambda_l^1 \quad \text{(cf. (4.4))}$$

it follows that the last two terms on the right side of (4.9) converge to zero uniformly in n as $m \to \infty$ through the sequence $\{m_k\}$. Now for each m the first term on the right converges to zero as $n \to \infty$ by the weak convergence of $P_n \to P$. Hence $M_t\phi$ is a $(P, \mathcal{A}_t)$ martingale, i.e., (3.10′) holds and so does (3.10). □

Observe that we have shown that

$$\bar{E}(\lambda_0^1) \subset \{P \in R^f(s, x)\colon (3.9'), (3.10'), (3.11) \text{ hold}\}.$$

If U is compact and

$$E(\lambda_0^1) = \{P \in R_c^f(s, x)\colon J_0^1(s, P) \leqq \lambda_0^1\},$$

then the above proof shows that $E(\lambda_0^1)$ is closed. In the proof that P satisfies (3.10′) we simply set $M_t^{1m} = M_t\phi$, $M_t^{2m} = 0$. We can even take $\lambda_0^1 = \infty$ and conclude that $R_c^f(s, x)$ is compact if U is.

THEOREM 4.7. *Assume* (3.5) *and that* J^2 *is continuous,* J^1 *is* l.s.c. *on* $\bar{E}(\lambda_0^1)$ *for all* $\lambda_0^1 \in [0, \infty)$. *If* $R^f(s, x)$ *is not empty then* $R'(s, x)$ *is not empty, i.e., there exists an optimal rule.*

Proof. If $\bar{P}$ is some element of $R^f(s, x)$, then we set

$$\lambda_0^1 = J_0^1(s, \bar{P}) < \infty.$$

Now $\bar{E}(\lambda_0^1)$ is a nonempty, compact subset of $R^f(s, x)$ and by (4.3) and the definition of $E(\lambda_0^1)$

$$\inf\{J_0^1(s, P): P \in R^f(s, x)\} = \inf\{J_0^1(s, P): P \in \bar{E}(\lambda_0^1)\}.$$

The result now follows since J_0^1 is l.s.c. □

COROLLARY 4.8. *Assume* (3.4), (3.5), *and the same continuity on J as in Theorem* 4.7. *If* $\tilde{\mathfrak{U}}_{sx}^f$ *is not empty, then there exists an optimal control that can be taken to be strict.*

This corollary follows directly from Theorem 4.7 and Remark 3.14.

Remark 4.9. We may ask whether there are any feasible controls, i.e., $\tilde{\mathfrak{U}}_{sx}^f \neq \varnothing$. In the presence of constraints this is a difficult question where notions of controllability come into play; however if we assume that there are no constraints, i.e., $m = n = 0$, then the question of feasibility reduces to (i) the existence of a control that (ii) has finite cost. But (i) holds if for some $\tilde{u}$ in U, $a(\cdot, \cdot, \tilde{u})$ and $b(\cdot, \cdot, \tilde{u})$ are bounded (cf. Stroock and Varadhan [19, Thm. 6.1.6]) or if $a(t, \cdot, \tilde{u})$ and $b(t, \cdot, \tilde{u})$ are locally Lipschitz continuous (cf. Métivier [17, Thm. 34.7]). As for (ii), i.e., whether the cost corresponding to the control process $u_t = \tilde{u}$ is finite, this is implied by any one of (in addition to the standing assumptions)

(α) $h_0(t, x) \leqq \kappa$, $f_0(t, x, \tilde{u}) \leqq \kappa e^{-\delta t}$, $\kappa < \infty$, $\delta > 0$;

(β) $h_0(t, x) \leqq \kappa$, $f_0(t, x, \tilde{u}) \leqq \kappa e^{-\delta t}(1 + |x|^q)$, $\kappa < \infty$, $\delta > 0$, $q \geqq 0$, (a, b) bounded;

(γ) $f_0(t, x, \tilde{u}) \leqq \kappa(1 + |x|^q)$, $\kappa < \infty$, $q \geqq 0$, $h_0(t, x) \leqq \kappa$ if $t \leqq T$, $h_0(t, x) = +\infty$ if $t > T$ for some T fixed.

4.10. The deterministic case. Let us now display some implications of our result in the deterministic case. We are given the following:

- U, a closed, σ-compact subset of a Banach space;
- B, a closed subset of $\bar{\mathbb{R}}_+ \times \mathbb{R}^d$;
- A, a measurable subset of $\mathbb{R}_+ \times \mathbb{R}^d \times U$ such that for each t, A_t, the t-section of A, is closed;
- b, a measurable function: $A \to \mathbb{R}^d$, continuous in (x, u) for each t such that for some $p \geqq 0$

$$|b(t, x, u)| \leqq k(1 + |x| + |u|^p);$$

- f, a measurable function: $A \to \mathbb{R}_+$, l.s.c. in (x, u) for each t, such that there exists a function $\tilde{f}$ in $C(U; \mathbb{R}_+)$ such that for all (t, x, u) in A

$$f(t, x, u) \geqq \tilde{f}(u), \qquad \lim_{\substack{|u| \to \infty \\ u \in U}} \frac{\tilde{f}(u)}{|u|^p} = +\infty;$$

- h, a function $B \to \mathbb{R}_+$, l.s.c.;
- b and f are such that for each (t, x) $\bigcup_{u \in U} \{b(t, x, u)\} \times [f(t, x, u), \infty)$ is convex;
- (s, x), a point in $\mathbb{R}_+ \times \mathbb{R}^d$ such that $(\{(s, x)\} \times U) \cap A \neq \varnothing$.

We write $AC(I; \mathbb{R}^d)$ for the set of absolutely continuous functions $I \to \mathbb{R}^d$. We can now define a "control."

A triple $\alpha = (X, u, S)$ is a *control*, i.e., $\alpha \in \mathcal{U}_{sx}$, if $S \in [s, \infty]$, $X \in AC([s, S]; \mathbb{R}^d)$, $u \in L^p_{\text{loc}}([s, S]; U)$ and for all $t \in [s, S]$

$$X(t) = x + \int_s^t b(\theta, X(\theta), u(\theta))\, d\theta.$$

The control α is *admissible* if $\alpha \in \mathcal{U}_{sx}$, $(t, X(t), u(t)) \in A$ a.e. t, $(S, X(S)) \in B$.
We consider three cases:

(I) $B \subset [0, M] \times \mathbb{R}^d$ for some $M < \infty$,

$$J(s, \alpha) := \int_s^S f(\theta, X(\theta), u(\theta))\, d\theta + h(S, X(S));$$

(II) $B = \{\infty\} \times \mathbb{R}^d$, $h = 0$

$$J(s, \alpha) := \int_s^\infty f(\theta, X(\theta), u(\theta))\, d\theta;$$

(III) $B = \bar{\mathbb{R}}_+ \times \mathbb{R}^d$ and

(a) h bounded

$$J(s, \alpha) := \int_s^S f(\theta, X(\theta), u(\theta))\, d\theta + e^{-\delta(S-s)} h(S, X(S)),$$

or

(b) $\lim_{t \to \infty} \inf_{x \in \mathbb{R}^d} h(t, x) = +\infty$

$$J(s, \alpha) := \begin{cases} \displaystyle\int_s^S f(\theta, X(\theta), u(\theta))\, d\theta + h(S, X(S)) & \text{if } S < \infty, \\ +\infty & \text{if } S = +\infty. \end{cases}$$

We say that α is a *feasible* control, i.e., $\alpha \in \mathcal{U}^f_{sx}$, if $\alpha \in \mathcal{U}_{sx}$, α is admissible and $J(s, \alpha) < \infty$. The control problem is

$$\inf \{J(s, \alpha) : \alpha \in \mathcal{U}^f_{sx}\}.$$

Observe that in case (I) the controller runs the process until the time S, chosen by him, but the form of the target set implies that $S \leqq M$. If in this case $A = \mathbb{R}_+ \times \mathbb{R}^d \times U$ and $B = \{T\} \times \mathbb{R}^d$, i.e., there are no state constraints, then $\mathcal{U}^f_{sx} \neq \varnothing$ provided there exist $\tilde{u} \in U$ such that

(4.10) $b(\cdot, \cdot, \tilde{u})$ is bounded, or

(4.11) $b(t, \cdot, \tilde{u})$ is Lipschitz continuous and $f(\cdot, \cdot, \tilde{u}) \in L^1_{\text{loc}}(\mathbb{R}_+ \times \mathbb{R}^d; \mathbb{R})$.

In case (II) the controller must run the process for all time. Here again if $A = \mathbb{R}_+ \times \mathbb{R}^d \times U$, i.e., there are no state constraints, then $\mathcal{U}^f_{sx} \neq \varnothing$ if there exists $\tilde{u} \in U$ such that either (4.10) holds and for some $\kappa < \infty$, $q < \infty$, $\delta > 0$,

$$f(t, x, \tilde{u}) \leqq \kappa\, e^{-\delta t}(1 + |x|^q),$$

or (4.11) holds and

$$f(t, x, \tilde{u}) \leqq \kappa\, e^{-\delta t}.$$

In case (III) the controller runs the process until the time S of his choice. In case (a) he may choose $S = +\infty$, i.e., never stop, and then the penalty for stopping is zero.

In case (b) he will never choose $S=+\infty$. In either case $\mathcal{U}_{sx}^{f}\neq\varnothing$ because any α with $S=s$ gives a feasible control.

THEOREM 4.11. *In cases* (I)-(III), *if* $\mathcal{U}_{sx}^{f}\neq\varnothing$ *then there exists an optimal control.*

Proof. This follows from Remark 2.6 and Corollary 4.8 if we define f_0 and h_0 appropriately. For a set Λ define $\chi_\Lambda(y)$ by

$$\chi_\Lambda(y)=\begin{cases}0 & \text{if } y\in\Lambda,\\ +\infty & \text{otherwise.}\end{cases}$$

Then χ_Λ is l.s.c. if Λ is closed. We set $D=\mathbb{R}_+\times\mathbb{R}^d$ so $\bar{D}_\infty=\{\infty\}\times\bar{\mathbb{R}}^d$. Extend $f(\cdot,\cdot,u)$ and $h(\cdot,\cdot)$ to D by setting them equal to zero where they are not already defined. Then define

$$f_0(t,x,u)=f(t,x,u)+\chi_A(t,x,u)$$

and in cases (I) and (II)

$$h_0(t,x)=h(t,x)+\chi_B(t,x).$$

Observe that in case (I) $h_0=+\infty$ on $\bar{D}_\infty$ and in case (II) $h_0=0$ on $\bar{D}_\infty$. In case (III) set

$$h_0(t,x)=\begin{cases}\exp[-\delta(t-s)]h(t,x) & \text{case (a)},\\ h(t,x) & \text{case (b)}, \quad t<\infty,\\ +\infty & \text{case (b)}, \quad t=+\infty.\end{cases}$$

Then $h_0=0$ on $\bar{D}_\infty$ in case (a) and $h_0=+\infty$ in case (b). Moreover, $f_0(t,\cdot,\cdot)$ is l.s.c. on $\mathbb{R}^d\times U$ and $h_0(\cdot,\cdot)$ is l.s.c. on $\bar{\bar{D}}$. □

Remark 4.12. Sometimes it is convenient to assume D to be only measurable rather than open. Then ρ is, in general, *not* a stopping time, but we can still obtain our result. Given any measure P on $\mathscr{C}$ let $\mathscr{C}^P$ be the completion of $\mathscr{C}$ under P and let

$$\tilde{\mathscr{C}}=\bigcap_P \mathscr{C}^P,$$

i.e., $\tilde{\mathscr{C}}$ is the universal completion of $\mathscr{C}$. Let $\mathscr{C}_t^P$ be the σ-algebra generated by $\mathscr{C}_t$ and the null sets of $\mathscr{C}^P$ and let

$$\tilde{\mathscr{C}}_t=\bigcap_P \mathscr{C}_t^P.$$

Note that $\tilde{\mathscr{C}}_t$ may be larger than the universal completion of $\mathscr{C}_t$! In any case $\{\tilde{\mathscr{C}}_t\}$ is a filtration and

$$\mathscr{C}_t\subset\tilde{\mathscr{C}}_t\subset\tilde{\mathscr{C}}.$$

Moreover, if $B\in\tilde{\mathscr{C}}$ and $P(B)=0$ for all P then $B\in\tilde{\mathscr{C}}_t$, and given any P and $A\in\tilde{\mathscr{C}}_t$ there exists $A_P\in\mathscr{C}_t$ and $N_P\in\mathscr{C}^P$ such that $A=A_P\Delta N_P$, $P(N_P)=0$, where Δ denotes symmetric difference. If we set

$$\tilde{\mathscr{C}}_{t+}=\bigcap_{s>t}\tilde{\mathscr{C}}_s,\ \tilde{\Omega}_t=\tilde{\mathscr{C}}_{t+}\times\mathscr{V}_t\times\mathscr{Z}_t,$$

then ρ is a $\{\tilde{\Omega}_t\}$ stopping time and we can work with $\{\tilde{\Omega}_t\}$ to establish the existence of an optimal rule.

Let us add that we can refer all processes back to Ω_t. For example, if $\{u_t\}$ is $\{\tilde{\Omega}_t\}$ progressive, hence $\{\bar{\Omega}_t\}$ progressive ($\{\bar{\Omega}_t\}$ is the augmented filtration), then by Lemma A1 there exists an (Ω_t) progressive process $\{u_t^*\}$ such that $u_t^*(\omega)=u_t(\omega)$ a.e.—$dt\,dP$. Furthermore, if S is an $\{\tilde{\Omega}_t\}$ stopping time, then $\mathbb{1}_{\{\cdot\geqq S\}}$ is $\{\tilde{\Omega}_t\}$ optional. By Lemma 7 of [6, Appendix 1] it is indistinguishable from an $\{\Omega_{t+}\}$ optional process, i.e., there

exists a wide sense $\{\mathcal{A}_t\}$ stopping time S^* [i.e., $\{S^* < t\} \in \mathcal{A}_t$] such that $S = S^*$ a.s. Thus by relaxing our definition of a control slightly, in particular, by replacing "stopping time" by "wide sense stopping time" in (C_4), we still have each rule (on $\tilde{\mathcal{A}}$) generating a relaxed control (on $\mathcal{A}$).

5. Optimal feedback controls. Recall that if α is a natural control, then it is specified by a measure P on $C(\mathbb{R}_+; \mathbb{R}^d)$, by a progressively measurable function

$$u: \mathbb{R}_+ \times C(\mathbb{R}_+; \mathbb{R}^d) \to U,$$

and by a function $S: C(\mathbb{R}_+; \mathbb{R}^d) \to \mathbb{R}$ such that $\{S \leqq t\} \in \mathcal{C}_t$ for all t (cf. Definition 3.8).

DEFINITION 5.1. The natural control $\alpha \in \mathcal{N}_{sx}$ is a *Markov control* if the corresponding control process u satisfies

$$u(t, \xi) = V(t, \xi(t))$$

for some Borel measurable function $V: D \to U$, and S is the first exit time after s of $(t, \xi(t))$ from D', some measurable subset of D.

In deterministic control theory the Markov controls are called feedback controls. Observe that if α is Markov then $X_\cdot$ is a Markov process since it satisfies

$$dX_t = b(t, X_t, V(t, X_t))\, dt + a^{1/2}(t, X_t, V(t, X_t))\, dw_t \tag{5.1}$$

for some Brownian motion $\{w_t\}$ and some square root of a. This explains the term "Markov control."

Our aim in this section is to show that if there is an optimal control, i.e., if $R'(s, x) \neq \varnothing$ for all (s, x) in D, then for any $(s, x) \in D$ there exists an optimal control that is a Markov control with the same function V and set D' for all (s, x), i.e., there exists an optimal control law. The method of proof is an abstract version of dynamic programming due to Krylov. Since it is a form of dynamic programming, it cannot accommodate soft constraints and so we assume

$$F_i = \lambda_i^1 = 0, \qquad i = 1, 2, \cdots, m, \quad (\text{or } m = 0),$$

$$G_i = \lambda_i^2 = 0, \qquad i = 1, 2, \cdots, n, \quad (\text{or } n = 0).$$

Since there are no constraints we write F for F_0 and, similarly, f, h, J for f_0, h_0, J_0. Let us also simplify the hard constraints: we assume that $f(t, x, u) < \infty$ if and only if $(x, u) \in A \times U(t)$ where A and $U(t)$ are closed and $U(t) \subset U$ for all t. Hence if $f(t, \cdot, \cdot)$ is l.s.c. on $A \times U(t)$, then it is l.s.c. on $\mathbb{R}^d \times U$. Let us also assume that $u^0 \in \bigcap_t U(t)$. If this last set is empty we can replace u^0 by u_t^0 and δ^0 by δ_t^0 (Dirac measure at u_t^0) with $u_t^0 \in U(t)$ and $u_\cdot^0 \in L^\infty_{\text{loc}}(\mathbb{R}_+; U)$. The results of the previous section as well as those of this section can easily be extended to this case.

It follows that if $(s, x) \in D$ and $x \in A$ then for all feasible controls $P\{X_t \in A, t \geqq s\} = 1$. For $(s, x) \in D$ but $x \notin A$ either $J = +\infty$ and the problem has no solution, or $S = s$ and $J = h(s, x)$ so the problem has a unique solution: $S = s$ (and the control process u is irrelevant) and $X_t = x$ provided $h(s, x) < \infty$. Hence the only case of importance is the one where $x \in A$. We write

$$D_A = \{(s, x) \in D: x \in A\}.$$

It is actually possible to allow constraints of the form $x(t) \in A(t)$ if $A(t)$ is closed, the graph of $A(\cdot)$ is measurable and $A(t) \subset A(t')$ if $t \geqq t'$, but we will not consider this extension.

Krylov's Markovian Selection Theorem was used by Stroock and Varadhan [19, § 12.2] to show that if existence (but not uniqueness) obtains in the martingale problem

corresponding to a differential operator of diffusion type, then it is always possible to choose a Markov solution. It was then used by Haussmann [11] and by El Karoui, Huu Nguyen, and Jeanblanc-Picqué [8] to show that there exist optimal controls if the data are bounded and the control set is compact.

Let us now define a *compact* subset $\tilde{R}(s, x)$ of $R^f(s, x)$ such that $(s, x) \to \tilde{R}(s, x)$ is measurable and

$$\inf\{J(s, P): P \in R(s, x)\} = \inf\{J(s, P): P \in \tilde{R}(s, x)\}.$$

We assume

(5.2) There exists a locally bounded, upper semicontinuous function λ defined on D_A such that for each $(s, x) \in D_A$

(i) $\tilde{R}(s, x) := \{P \in R_0(s, x): J(s, P) \leqq \lambda(s, x)\} \neq \varnothing$,

(ii) $\rho(\cdot)$ is continuous P-a.s. for all $P \in R^f(s, x)$.

We can extend R and $\tilde{R}$ to $\bar{D}_A$ by setting

$$\tilde{R}(s, x) = R(s, x) = \{\delta_x \times \delta^* \times \delta_{\mathbb{1}_{\{\cdot \geqq s\}}}\}$$

for $(s, x) \in \partial D \cap \bar{D}_A$.

If U is compact, then we can dispense with the continuity of ρ in (5.2). We simply define $\tilde{R}(s, x) = R_c^f(s, x)$ for all $(s, x) \in \mathbb{R}_+ \times \mathbb{R}^d$ and assume that it is nonempty. Recall that $R_c^f(s, x)$ is compact if U is. In future proofs we will not mention this case unless the proof is different from the one given.

Now assume

(5.3) $(s, x, P) \to J(s, P)$ is lower semicontinuous for $(s, x) \in \bar{D}_A$ and $P \in \tilde{R}(s, x)$.

We discuss (5.2) and (5.3) in Remark 5.3 and Lemma 5.4.

PROPOSITION 5.2. *Assume* (3.5), (5.2), *and* (5.3). *Then for* (s, x) *in* $\bar{D}_A$ *the sets* $\tilde{R}(s, x)$ *are nonvoid, compact and the map* $(s, x) \to \tilde{R}(s, x)$ *is a measurable map of* $\bar{D}_A$ *into* comp $(M_1(C \times V \times Z))$.

Proof. By (5.2), $\tilde{R}(s, x) \neq \varnothing$. Observe that $\tilde{R}(s, x) = E(\lambda(s, x))$ of Proposition 4.6, hence Propositions 4.5 and 4.6 imply that $\bar{E}(\lambda(s, x))$ is compact. Since $E(\lambda(s, x)) = i_0 \bar{E}(\lambda(s, x))$ and i_0 is continuous (since ρ is), then $\tilde{R}(s, x)$ is compact.

To establish measurability we use a result of Stroock and Varadhan [19, Lemma 12.1.8]. We must show that if $(s_n, x_n) \to (s, x)$, $P_n \in \tilde{R}(s_n, x_n)$, then $\{P_n\}$ has a limit point in $\tilde{R}(s, x)$. We will show that $\{P_n\}$ is tight by showing that (A2) and (A4) hold with $\Lambda = \overline{\cup_n \{(s_n, x_n)\}}$. But, if $p > 0$

$$P_n(\Sigma_M^c) \leqq P_n\{\|\xi\|_T \geqq M\} + P_n\left\{\int_0^T |\eta_t|^p \, dt \geqq M\right\} + P_n\left\{\int_0^T \tilde{f}(\eta_t) \, dt \geqq M\right\}$$

$$\leqq K(T)[\lambda(s_n, x_n) + |x_n| + 1]/M$$

(cf. the Appendix and the proof of Proposition 4.5), and the last expression goes to zero as $M \to \infty$, uniformly in n. The proof in the case $p = 0$ is similar. The proof that (A4) holds is identical to the one given in Proposition 4.5.

Hence there exists P in $M_1(C \times V \times Z)$ such that (for a subsequence) $P_n \to P$. We will now show that P is in $\tilde{R}(s, x)$; the proof is similar to that given in Proposition 4.6 so we only indicate where it must be modified. To establish (3.11′) set

$$B^m = \{\omega = (\xi, \eta, \zeta): \|\xi(\cdot) - x\|_{s-1/m} \leqq 1/m, \eta(t) = \delta^0 \text{ a.e.}, \zeta(t) = 0, 0 \leqq t \leqq s - 1/m\}.$$

Then B^m is closed, $B^{m+1} \subset B^m$, and

$$\bigcap_m B^m = B := \{\omega : \omega(t) = (x, \delta^0, 0), 0 \leqq t < s\}.$$

Since $(s_n, x_n) \to (s, x)$ and P_n satisfies (3.11'),

$$P(B^m) \geqq \limsup_{n\to\infty} P_n(B^m) = 1.$$

Since B^m decreases to B, $P(B) = 1$. Moreover, if

$$\tilde{B} := \{(\xi(t), \eta(t)) = (\xi(y), \delta^*): t > y\},$$

then $\bar{\tilde{B}} \subset \{\tilde{B} \cup D_\rho\}$ if D_ρ is the discontinuity set of ρ. Thus if ρ is P-a.s. continuous, then $P(\tilde{B}) = P(\bar{\tilde{B}}) \geqq \limsup P_n(\bar{\tilde{B}}) = 1$. Hence P satisfies (3.11') if ρ is P-a.s. continuous, and (3.11) otherwise. Recall the overbar denotes closure.

The proof of (3.9') is unchanged from that given in Proposition 4.6 with λ_l^1 replaced by $\sup_n \lambda(s_n, x_n)$. For (3.10') again the proof is unchanged; the only awkward point arises when $t = s < s_n$. Here we observe that

$$\phi(\xi_r) - \phi(\xi_s) - \int_s^T L\phi(\theta, \xi_\theta, \eta_\theta)\, d\theta = \left\{\phi(\xi_r) - \phi(\xi_s) - \int_{s_n}^T L\phi\, d\theta\right\} - \int_s^{s_n} L\phi\, d\theta.$$

The term inside braces is a P_n martingale (since $\xi_s = \xi_{s_n}$ P_n-a.s.) and

$$\int_s^{s_n} L\phi(\theta, \xi_\theta, \eta_\theta)\, d\theta = \int_s^{s_n} L\phi(\theta, x_n, u_0)\, d\theta \quad P_n\text{-a.s.}$$

The last integral tends to zero as $n \to \infty$. We conclude that P is in $R_0(s, x)$.

Finally, (5.2) and (5.3) imply

$$\begin{aligned} J(s, P) &\leqq \liminf_{n\to\infty} J(s_n, P_n) \\ &\leqq \liminf_{n\to\infty} \lambda(s_n, x_n) \\ &\leqq \lambda(s, x) \end{aligned}$$

so that P is in $\tilde{R}(s, x)$. This completes the proof. □

Remark 5.3. Assumption (5.2)(i) holds for the cases mentioned in Remark 4.9 for any $q < \infty$. Recall that the P-a.s. continuity of ρ was discussed in Remark 4.2.

The following lemma gives a sufficient condition for (5.3).

LEMMA 5.4. *For each* (s, x) *in* D_A *assume that*

(i) $\omega \to F(s, \omega)$ *is* l.s.c. P-a.s. *for* $P \in \tilde{R}(s, x)$,

(ii) *There exist* $\delta > 0$ *and a function* $\chi(t)$, *integrable on* $[s, s+\delta)$, *such that* $f(t, x', u^0) \leqq \chi(t)$ *for* $s \leqq t < s+\delta$, $|x - x'| < \delta$, $x' \in A$.

Then (5.3) *holds. Note that the lower semicontinuity of* F *is discussed in Remark* 4.3 *and Lemma* 4.4.

Proof. Assume that $(s_n, x_n, P_n) \to (s, x, P)$ with $(s_n, x_n) \in D_A$ and $P_n \in \tilde{R}(s_n, x_n)$. It suffices to consider two cases: $s_n \uparrow s$ or $s_n \downarrow s$. In the latter case (ii) implies that

$$\begin{aligned} P_n \int_s^{s_n} f(t, \xi_t, \eta_t)\, dt &= P_n \int_s^{s_n} f(t, x_n, u^0)\, dt \\ &\leqq \int_s^{s_n} \chi(t)\, dt \end{aligned}$$

if n is so large that $s_n - s < \delta$ and $|x_n - x| < \delta$. Hence

$$\limsup_{n\to\infty} P_n[F(s, \omega) - F(s_n, \omega)] = \limsup_{n\to\infty} P_n \int_s^{s_n} f(t, \xi_t, \eta_t)\, dt \leqq 0. \tag{5.4}$$

In the case where $s_n \uparrow s$, (5.4) also holds since $f \geqq 0$.

But now since $\omega \to F(s, \omega)$ is l.s.c. P-a.s., we have

$$PF(s,\omega) \leqq \liminf_n P_n F(s,\omega)$$
$$\leqq \liminf_n P_n F(s_n,\omega) + \limsup_n P_n[F(s,\omega) - F(s_n,\omega)]$$

so the result follows from (5.4). □

Let us now define the *value function*: for (s, x) in $\bar{D}_A$

$$v(s,x) := \inf\{J(s,P) : P \in R(s,x)\}.$$

As we will assume (5.2), we have

$$v(s,x) = \inf\{J(s,P) : P \in \tilde{R}(s,x)\},$$
$$R_0'(s,x) = \{P \in \tilde{R}(s,x) : J(s,P) = v(s,x)\}. \tag{5.5}$$

Let us define

$$D_A' = \{(s,x) \in D_A : v(s,x) < h(s,x)\},$$

hence $v(s, x) = h(s, x)$ for $(s, x) \in \bar{D}_A \backslash D_A'$.

Since $J(x, P) = h(s, x)$ if the controller stops immediately, and hence $v(s,x) \leqq h(s,x)$, it follows that for $(s,x) \in \bar{D}_A \backslash D_A'$ the controller can do no better than to stop immediately. The next result states that there is an optimal control that stops immediately if and only if $(s,x) \in \bar{D}_A \backslash D_A'$, and that such controls can be gotten as a *measurable* selector of R_0'.

If $P \in R(s,x)$, then $P(\cdot \mid \zeta(s))$ exists as a regular conditional probability distribution (r.c.p.d.); we denote it by $P_{\zeta(s)}$ and we write

$$P_z(\cdot) = P(\cdot \mid \zeta(s) = z), \qquad z = 0, 1.$$

Note that if $P \in R_0^f(s,x)$, then $P_{\zeta(s)} \in R_0^f(s,x)$ P-a.s. Indeed P is a convex combination of P_0 and P_1 so (3.11′) holds for $P_{\zeta(s)}$ P-a.s. since it holds for P. The same is true for (3.9′) and the feasibility condition $J(s,P) < \infty$ since only linear functionals of P are involved. Finally, if $\theta > t \geqq s$ and Ψ is an $\mathcal{A}_t$ measurable bounded function, then

$$P\{\zeta(s) = z\} P_z([M_\theta \phi - M_t \phi]\Psi) = P\{\mathbb{1}_{\{\zeta(s)=z\}}[M_\theta \phi - M_t \phi]\Psi\}$$
$$= 0$$

so that (3.10′) holds P-a.s.

LEMMA 5.5. *Assume* (3.5), (5.2), *and* (5.3). *Then* $v : \bar{D}_A \to \mathbb{R}_+$ *is measurable and there exists a measurable selector* H *of* R_0', *i.e.*, $H \in \operatorname{meas}(R_0')$, *such that for* $(s,x) \in \bar{D}_A$

$$H(s,x)\{\zeta(s) = 0\} = \mathbb{1}_{D_A'}(s,x). \tag{5.6}$$

Proof. According to Stroock and Varadhan [19, Lemma 12.1.7], the mapping $K \to \inf\{J(s,P) : P \in K\}$ is measurable for $K \in \operatorname{comp}(M_1(C \times V \times Z))$. The measurability of v now follows from Proposition 5.2.

From Lemma 12.1.7 of [19], Proposition 5.2, and (5.5), it also follows that $(s,x) \to R_0'(s,x)$ is measurable, hence Theorem 12.1.10 of [19] implies that there is a measurable selector Q of R_0'.

Let us set

$$Q_z(\cdot) = Q(\cdot \mid \zeta(s) = z), \qquad z = 0, 1,$$

the r.c.p.d. of Q given $\zeta(s) = z$. Then

$$Q = Q\{\zeta(s) = 1\} Q_1 + Q\{\zeta(s) = 0\} Q_0.$$

Observe that $J(s, Q_1) = h(s, x)$ so

$$v(s, x) \leqq h(s, x).$$

Moreover, if $(s, x) \in D'_A$ (so $v(s, x) < h(s, x)$) then

$$v(s, x) = J(s, Q) = Q\{\zeta(s) = 1\}h(s, x) + Q\{\zeta(s) = 0\}J(s, Q_0)$$

and hence $J(s, Q_0) < h(s, x)$. Now the optimality of Q implies that $Q\{\zeta(s) = 1\} = 0$ and $v(s, x) = J(s, Q_0)$. Conversely, if $(s, x) \notin D'_A$ (so $v(s, x) = h(s, x)$) then

$$h(s, x) = Q\{\zeta(s) = 1\}h(s, x) + Q\{\zeta(s) = 0\}J(s, Q_0),$$

so either $Q\{\zeta(s) = 0\} = 0$ or $J(s, Q_0) = h(s, x)$. In the latter case the controller may not stop immediately, i.e., he continues without accumulating any further cost, but then it is also optimal to stop immediately! Hence if we define

$$H(s, x) = \mathbb{1}_{D'_A}(s, x)Q(s, x) + \mathbb{1}_{\bar{D}_A \setminus D'_A}(s, x)Q_1(s, x),$$

then H is a measurable selector of $R'_0(s, x)$ that satisfies (5.6)—note that the map $(s, x) \to Q_1(s, x)$ is measurable by Lemma A.10. The result now follows. □

5.6. Notation. The Markovian Selection Theorem requires a certain structure that we now present. We follow Stroock and Varadhan [19, § 12.2]. Recall that $\{\mathcal{A}_t\}$ is the canonic Borel filtration on $\Omega = C \times V \times Z$. Since Ω is a Polish space and since each $\mathcal{A}_t$ is countably generated, regular conditional probability distributions (r.p.c.d.) given $\mathcal{A}_t$ exist for any P in $M_1(\Omega)$ (cf. [19] for the definition). If P is in $M_1(\Omega)$ and τ is a finite stopping time we denote a r.c.p.d. of P given $\mathcal{A}_\tau$ by P^τ_ω. It exists since $\mathcal{A}_\tau$ is countably generated. Note that we *cannot* replace $\{\mathcal{A}_\tau\}$ by the usual augmented filtration $\{\bar{\mathcal{A}}_t\}$ since $\bar{\mathcal{A}}_\tau$ is not countably generated.

For a finite stopping time τ we say that Q_ω is a *τ-transition probability* (or τ-t.p.) if Q_ω is in $M_1(\Omega)$ for each ω and

(i) $\omega \to Q_\omega(A)$ is $\mathcal{A}_\tau$ measurable for A in $\mathcal{A}$,

(ii) $Q_\omega\{(\bar{\xi}, \bar{\eta}, \bar{\zeta}) : (\bar{\xi}(\tau(\omega)), \bar{\zeta}(\tau(\omega))) = (\xi(\tau(\omega)), \zeta(\tau(\omega)))\} = 1$

for each $\omega = (\xi, \eta, \zeta)$ in Ω.

For each ω fixed with $\tau(\omega) < \infty$ we define $\mathcal{A}^{\tau(\omega)}$ to be the σ-algebra generated by

$$\left\{(\xi', \eta', \zeta') \in \Omega : \xi'_t \in \tilde{A}, \int_{\tau(\omega)}^{t} \eta'_\theta \, d\theta \in B, \zeta'_t = 0\right\}$$

for any $t \geqq \tau(\omega)$, $\tilde{A}$ and B Borel sets in $\mathbb{R}^d$ and $M_1(U)$, respectively. These are the events that take place after the fixed time $\tau(\omega)$. Given a finite stopping time τ, a τ-t.p. Q_ω and $\omega' \in \Omega$, then according to Lemma A.11 for any ω such that

$$(\xi(\tau(\omega)), \zeta(\tau(\omega))) = (\xi'(\tau(\omega)), \zeta'(\tau(\omega)))$$

there exists a unique measure $\delta_{\omega'} / \tau / Q_\omega \in M_1(\Omega)$ such that

$$\begin{aligned} &\delta_{\omega'} / \tau / Q_\omega\{\bar{\omega} : \bar{\omega}(t) = \omega'(t), t \leqq \tau(\omega)\} = 1, \\ &\delta_{\omega'} / \tau / Q_\omega(\tilde{A}) = Q_\omega(\tilde{A}), \qquad \tilde{A} \in \mathcal{A}^{\tau(\omega)}. \end{aligned} \tag{5.7}$$

Note that when we write $\bar{\omega}(t) = \omega'(t)$, $t \leqq \tau(\omega)$, we mean $(\bar{\xi}(t), \bar{\zeta}(t)) = (\xi'(t)), \zeta'(t)$ for all $t \leqq \tau(\omega)$ and $\bar{\eta}(t) = \eta'(t)$ for almost all $t \leqq \tau(\omega)$. In case $\omega' = \omega$ we write

$$\delta_\omega / \tau / Q_\omega = Q^\tau_\omega;$$

moreover, if $\omega' = (\xi(\tau(\omega)), \delta^0, \zeta \mathbb{1}_{\{\cdot \geqq \tau(\omega)\}})$ then we write

$$\delta_{\omega'} / \tau / Q_\omega = \mathring{Q}^\tau_\omega.$$

In addition, the proof of Theorem 6.1.2 of [19], shows that given P in $M_1(\Omega)$ and a τ-t.p. Q_ω there exists a unique element $P/\tau/Q$ in $M_1(\Omega)$ such that

$$\text{(5.8)} \quad \begin{aligned} &\text{(i)} \quad P/\tau/Q(\tilde{A}) = P(\tilde{A}) \text{ for } \tilde{A} \in \mathcal{A}_\tau, \\ &\text{(ii)} \quad Q^\tau_\omega \text{ is a r.c.p.d. of } P/\tau/Q \text{ given } \mathcal{A}_\tau. \end{aligned}$$

From (5.7) it follows that *on* $\mathcal{A}^{\tau(\omega)}$, $\mathring{Q}^\tau_\omega = Q_\omega = Q^\tau_\omega$ is a r.c.p.d. of $P/\tau/Q$ given $\mathcal{A}_\tau$. On the other hand, if P^τ_ω is a r.c.p.d. of P given $\mathcal{A}_\tau$ then $Q_\omega := P^\tau_\omega$ is a τ-t.p. and $P/\tau/Q = P$, $Q^\tau_\omega = P^\tau_\omega = Q_\omega$ on Ω (not just on $\mathcal{A}^{\tau(\omega)}$).

Recall that

$$F(t, \omega) = \int_t^{y(\omega)} f(\theta, \xi(\theta), \eta(\theta))\, d\theta + h(y(\omega), \xi(y(\omega))).$$

In this definition y is the first time *after* t at which $(\theta, \xi(\theta))$ leaves D or ζ jumps to 1, i.e., $y(t, \omega) := \inf\{\theta \geqq t : (\theta, \xi(\theta)) \notin D \text{ or } \zeta(\theta) = 1\}$. From now on this dependence of y on t is important so we write $y(t, \omega)$. If we do not wish to emphasize ω we write $y(t)$ for $y(t, \omega)$. Observe now that

$$\omega' \to F(\tau(\omega), \omega') = \int_{\tau(\omega)}^{y(\tau(\omega), \omega')} f(t, \xi'(\tau), \eta'(t))\, dt + h(y(\omega'), \xi(y(\omega')))$$

is $\mathcal{A}^{\tau(\omega)}$ measurable so

$$Q_\omega F(\tau(\omega), \cdot) = Q^\tau_\omega F(\tau(\omega), \cdot) = \mathring{Q}^\tau_\omega F(\tau(\omega), \cdot). \tag{5.9}$$

The same is true if we replace F by $\int_{\tau(\omega)}^{T \wedge y} |\eta'_t|^p\, dt$ with $T \geqq \tau(\omega)$.

If $P \in R^f(s, x)$, $\tau \in [s, y(s)]$ is a finite stopping time and H is as in Lemma 5.5, then by (5.6) for $(s, x) \in \bar{D}_A$

$$H(s, x) = \mathbb{1}_{D_A}(s, x) H(s, x, 0) + \mathbb{1}_{\bar{D}_A \setminus D_A}(s, x) H(s, x, 1) \tag{5.10}$$

where $H(s, x, z) := H_z(s, x)$ is the r.c.p.d. of $H(s, x)$ given $\zeta(s) = z$. Now set $Q_\omega = H(\tau(\omega), \xi(\tau(\omega)), \zeta(\tau(\omega)))$. Then Q_ω is a τ-t.p. (cf. Lemma A.10) and $Q_\omega = \mathring{Q}^\tau_\omega$. We write $P/\tau/H$ for $P/\tau/Q$. From (5.6) and (5.10) we have

$$\begin{aligned} J(\tau, H(\tau, \xi(\tau))) &= \mathbb{1}_{D_A}(\tau, \xi(\tau)) J(\tau, H(\tau, \xi(\tau), \zeta(\tau))) \\ &\quad + \mathbb{1}_{\bar{D}_A \setminus D_A}(\tau, \xi(\tau)) J(\tau, H(\tau, \xi(\tau), \zeta(\tau))) \\ &= J(\tau, H(\tau, \xi(\tau), \zeta(\tau))) \quad P\text{-a.s.} \end{aligned}$$

and now from (5.9)

$$PJ(\tau, H(\tau, \xi(\tau))) = J(\tau, P/\tau/H). \tag{5.11}$$

The next two results are central to the argument that follows. They fail when soft constraints are present—for this reason we assumed that there are none. The first result states that a feasible control remains a feasible control for problems starting at a later time from a point reached at that time. The second result says (more or less) that if we take a feasible control and at some later time switch to a control that is optimal *from then on*, then this concatenated object is still a feasible control.

LEMMA 5.7 (closure under conditioning). *If* $P \in R^f(s, x)$ *and if* $\tau \in [s, y(s)]$ *is a finite stopping time, then there exists a null set* N *in* $\mathcal{A}_\tau$ *such that* $\mathring{P}^\tau_\omega \in R^f(\tau(\omega), \xi(\tau(\omega)))$ *for* $\omega \notin N$.

Proof. By definition $\mathring{P}^\tau_\omega = \delta_{\omega'}/\tau/P^\tau_\omega$ where $\omega = (\xi, \eta, \zeta)$, $\omega' = (\xi(\tau(\omega)), \delta^0, \mathbb{1}_{\{\cdot \geqq \tau(\omega)\}})$, and P^τ_ω is a r.c.p.d. of P given $\mathcal{A}_\tau$, so $\mathring{P}^\tau_\omega$ satisfies (3.11). Moreover, as in the proof of Theorem 6.1.3 of [19], P^τ_ω inherits the martingale property from P for ω not

in some $\mathcal{A}_\tau$-null set N_0, and as in the proof of Theorem 6.2.1 of [19], $\mathring{P}^\tau_\omega$ inherits the martingale property from P^τ_ω. Hence (3.10) holds for $\mathring{P}^\tau_\omega$. Next

$$\mathring{P}^\tau_\omega \int_0^{T\wedge y(\tau(\omega))} |\eta|^p\,dt = \mathbb{1}_{\{\tau(\omega)\geqq T\}} T|u^0|^p + \mathbb{1}_{\{\tau(\omega)<T\}}\left[\tau(\omega)|u^0|^p + P^\tau_\omega \int_{\tau(\omega)}^{T\wedge y(\tau(\omega))} |\eta|^p\,dt\right]$$

since $\int_{\tau(\omega)}^{T\wedge y(\tau(\omega))} |\eta|^p\,dt$ is $\mathcal{A}^{\tau(\omega)}$ measurable if $T>\tau(\omega)$. But for any $k>0$

$$PP^\tau_\omega \int_0^k |\eta|^p\,dt = P\int_0^k |\eta|^p\,dt < \infty$$

so for ω not in an $\mathcal{A}_\tau$-null set N_k, $k=1,2,\cdots$

$$P^\tau_\omega \int_0^k |\eta|^p\,dt < \infty$$

and hence $\mathring{P}^\tau_\omega$ is in $R(\tau(\omega),\xi(\tau(\omega)))$ if $\omega\notin \bar{N}=\cup_{i=0}^\infty N_i$. A similar argument with $F(\tau(\omega),\cdot)$ replacing $\int_{\tau(\omega)}^T |\eta|^p\,dt$ produces a null set $\tilde{N}$ so that $\mathring{P}^\tau_\omega$ is feasible for $\omega\notin N=\bar{N}\cup\tilde{N}$. □

LEMMA 5.8 (closure under concatenation). *Assume* (5.2). *If* $P\in R^f(s,x)$, *if* H *is as in Lemma* 5.5 *and if* $\tau\in[s,y(s)]$ *is a finite stopping time, then*

$$P/\tau/H\in R^f(s,x).$$

Proof. Write $\tilde{P}$ for $P/\tau/H$. By (5.8)(i) it follows that $\tilde{P}$ inherits the property (3.11) from P. Moreover, (3.10) for $\tilde{P}$ follows from Theorem 1.2.10 of [19]. Finally, if $Q_\omega=H(\tau(\omega),\xi(\tau(\omega)),\zeta(\tau(\omega)))$, then

$$\tilde{P}\int_0^{T\wedge y(s)} |\eta|^p\,dt = P\left(\mathbb{1}_{\{T\wedge y\leqq\tau\}}\int_0^{T\wedge y(s)} |\eta|^p\,dt\right) + P\left\{\mathbb{1}_{\{T\wedge y(s)>\tau\}} Q^\tau_\omega \int_0^{T\wedge y(s)} |\eta|^p\,dt\right\} < \infty.$$

Indeed if $\tau<T\wedge y(s)$ then $y(s)=y(\tau(\omega))$ and for $T>\tau(\omega)$ and m sufficiently large

$$\begin{aligned}
Q^\tau_\omega \int_0^{T\wedge y(s)} |\eta|^p\,dt &= \int_0^{\tau(\omega)} |\eta|^p\,dt + \mathbb{1}_{D'_A}(\tau,\xi(\tau))H(\tau,\xi(\tau),0)\int_{\tau(\omega)}^{T\wedge y(s)} |\eta|^p\,dt\\
&\leqq \int_0^{\tau(\omega)} |\eta|^p\,dt + \nu_m \mathbb{1}_{D'_A}(\tau,\xi(\tau))H(\tau,\xi(\tau),0)\\
&\quad\times\int_{\tau(\omega)}^{T\wedge y(s)} \tilde{f}\,dt + [T-\tau(\omega)]m^p\\
&\leqq \int_0^{\tau(\omega)} |\eta|^p\,dt + \nu_m v(\tau(\omega),\xi(\tau(\omega))) + Tm^p\\
&\leqq \int_0^{\tau(\omega)} |\eta|^p\,dt + \nu_m J(\tau(\omega),\mathring{P}^\tau_\omega) + Tm^p \quad \text{a.s.}\\
&\leqq \int_0^{\tau(\omega)} |\eta|^p\,dt + \nu_m J(s,\mathring{P}^\tau_\omega) + Tm^p
\end{aligned}$$

where the second inequality holds because for $(\tau,\xi(\tau))\in D'_A$, $H(\tau,\xi(\tau),0)=H(\tau,\xi(\tau))\in R'_0(\tau,\xi(\tau))$. The third inequality follows from Lemma 5.7. Hence

$$\tilde{P}\int_0^{T\wedge y(s)} |\eta|^p\,dt \leqq P\int_0^{T\wedge y(s)} |\eta|^p\,ds + \nu_m J(s,P) + Tm^p < \infty.$$

Similarly, we can show that

$$J(s, \tilde{P}) = P\left[\int_s^\tau f(t, \xi(t), \eta(t))\, dt + \mathbb{1}_{\bar{D}_A \setminus D_A'}(\tau, \xi(\tau))h(\tau, \xi(\tau)) + \mathbb{1}_{D_A'}(\tau, \xi(\tau))v(\tau, \xi(\tau))\right]$$
$$\leqq J(s, P) < \infty$$

since $P \in R^f(s, x)$. □

We observe that if in Lemma 5.7, $P \in R_0^f(s, x)$ or $R_c^f(s, x)$ then $\mathring{P}_\omega^\tau \in R_0^f(\tau(\omega), \xi(\tau(\omega)))$ or $R_c^f(\tau(\omega), \xi(\tau(\omega)))$ as the case may be. Moreover, in Lemma 5.8, $P/\tau/H \in R_0^f(s, x)$ since $H \in R_0^f(s, x)$.

Let us now establish the dynamic programming result. For any function ϕ set

$$\Gamma_t(s, \phi) = \int_s^t f(\theta, \xi(\theta), \eta(\theta))\, d\theta + \phi(t, \xi(t)),$$

$$\Gamma_t^*(s, \phi) = \Gamma_{t \wedge y(s)}(s, \phi).$$

PROPOSITION 5.9. *Assume* (3.5), (5.2), *and* (5.3). *Then*

(5.12) (a) *If* $\tau \in [s, y(s)]$ *is a finite stopping time, then*

$$v(s, x) = \inf\{P\Gamma_\tau^*(s, v) : P \in R^f(s, x)\};$$

(b) *For any* $P \in R^f(s, x)$, $\Gamma_\cdot^*(s, v)$ *is a* $(P, \mathcal{A}_t)$ *submartingale*;

(c) $P \in R'(s, x)$ *if and only if* $\Gamma_\cdot^*(s, v)$ *is a* $(P, \mathcal{A}_t)$ *martingale and* $(y(s), \xi(y(s))) \notin D_A'$ P-a.s.

Proof. For any $P \in R^f(s, x)$ let $\mu(dt, dx')$ be the distribution of $(\tau, \xi(\tau))$ under P. Note that μ has support in $\bar{D}_A$. With H as in Lemma 5.5, we have

$$\begin{aligned} Pv(\tau, \xi(\tau)) &= \int_{\bar{D}_A} v(t, x')\mu(dt, dx') \\ &= \int_{\bar{D}_A} J(t, H(t, x'))\mu(dt, dx') \\ &= PJ(\tau, H(\tau, \xi(\tau))) \\ &= J(\tau, P/\tau/H), \end{aligned}$$

where we have used (5.11). The above and Lemma 5.8 imply

$$\begin{aligned} P\Gamma_\tau(s, v) &= P\int_s^\tau f(\theta, \xi(\theta), \mu(\theta))\, d\theta + J(\tau, P/\tau/H) \\ &= J(s, P/\tau/H) \qquad (5.13) \\ &\geqq \inf\{J(s, Q) : Q \in R^f(s, x)\} \\ &= v(s, x). \end{aligned}$$

On the other hand, by Lemma 5.7 and (5.9)

$$Pv(\tau, \xi(\tau)) \leqq PJ(\tau, \mathring{P}_\omega^\tau) = PJ(\tau, P_\omega^\tau) = J(\tau, P),$$

so from (5.13) and the definition of Γ

$$\begin{aligned} v(s, x) &\leqq P\Gamma_\tau(s, v) \\ &\leqq P\int_s^\tau f(t, \xi(t), \eta(t))\, dt + J(\tau, P) \\ &= J(s, P). \end{aligned}$$

Property (5.12)(a) follows by taking inf over $P \in R^f(s, x)$ in the last string of inequalities.

To show that $\Gamma^*_\cdot(s, v)$ is a submartingale, it suffices to take $t < y(s)$ and to observe that

$$P\{\Gamma^*_{t+h}(s, v) - \Gamma^*_t(s, v) \mid \mathcal{A}_t\} = P\{\Gamma^*_{t+h}(t, v) \mid \mathcal{A}_t\} - v(t, \xi(t))$$
$$= \mathring{P}^t_\omega \Gamma^*_{t+h}(t, v) - v(t, \xi(t))$$

since $\Gamma^*_{t+h}(t, v)$ is $\mathcal{A}^t$-measurable. Recall that $t < y(s)$ implies that $y(t) = y(s)$. By Lemma 5.7, $\mathring{P}^t_\omega \in R^f(t, \xi(t))$ P-a.s. so by (5.13) with $s = t$ and $\tau = (t+h) \wedge y$

$$\mathring{P}^t_\omega \Gamma^*_{t+h}(t, v) \geqq v(t, \xi(t)) \quad P\text{-a.s.}$$

hence Γ^* is a submartingale.

If Γ^* is a martingale and $(y, \xi(y)) \notin D'_A$, i.e., $v(y, \xi(y)) = h(y, \xi(y))$, then

$$v(s, x) = P\Gamma^*_s(s, v) = P\Gamma^*_{y(s)}(s, v) = J(s, P).$$

Thus $P \in R'(s, x)$.

Conversely, if $P \in R'(s, x)$ then by (a) with $\tau = y(s) \wedge t$

$$v(s, x) \leqq P\Gamma_\tau(s, v)$$
$$\leqq P \int_s^\tau f(\theta, \xi(\theta), \eta(\theta))\, d\theta + PJ(\tau, P^\tau_\omega)$$
$$= J(s, P)$$
$$= v(s, x)$$

where (5.9) and Lemma 5.7 were used to obtain the second inequality. Hence $\Gamma^*_\cdot(s, v)$ is a submartingale with constant mean, hence a martingale.

Moreover by (b) and the optimality of P

$$P\Gamma_{y(s)}(s, v) \geqq P\Gamma_s(s, v) = v(s, x) = P\Gamma_{y(s)}(s, h)$$

so that $Pv(y, \xi(y)) \geqq Ph(y, \xi(y))$. Since $v(\cdot, \cdot) \leqq h(\cdot, \cdot)$ then $v(y, \xi(y)) = h(y, \xi(y))$ P-a.s. □

Remark 5.10. The above proposition allows us to eliminate S (or ζ) from further consideration. Indeed if we set

$$D' := \{(s, x) \in D\colon v(s, x) < h(s, x)\},$$

then $D'_A = D' \cap (\mathbb{R}_+ \times A)$. We call the reduced problem the control problem with D replaced by D'. Let us write v' for the corresponding value function. Assume that $P \in R'(s, x)$. Let ρ' be the first exist time of $(t, \xi(t))$ from D' (we assume D' is open; cf. below). Then Proposition 5.9(c) implies that

$$v(s, x) = P\left\{\int_s^{\rho' \wedge \Delta} f(t, \xi(t), \eta(t))\, dt + h(\rho' \wedge \Delta, \xi(\rho' \wedge \Delta))\right\}.$$

Hence if $P' = P/\rho'/H$, i.e., P' is P except that the stopping time S is replaced by $S' = S \wedge \rho'$, then $v(s, x) \geqq J(s, P') \geqq v'(s, x)$. But any feasible control for the reduced problem is feasible for the original, so $v(s, x) = v'(s, x)$ and P' is optimal for the reduced problem. Conversely, if the law of the canonic relaxed control (X_t, μ_t, S') is optimal for the reduced problem, then that of $(X_t, \mu_t, S' \wedge \rho')$ is optimal for the original by Proposition 5.9(c).

Finally, we show that if P is optimal for the reduced problem then $P\{\Delta(\zeta) < \rho'\} = 0$. Assume $P\{\Delta(\zeta) < \rho'\} > 0$. Since $v \leqq h$ with strict inequality on D',

$$\begin{aligned} v'(s, x) &= J(s, P) \\ &= P\left\{ \mathbb{1}_{\{\Delta < \rho'\}} \left[\int_s^{\Delta} f\,dt + h(\Delta, \xi(\Delta)) \right] + \mathbb{1}_{\{\Delta = \rho'\}} \left[\int_s^{\rho'} f\,dt + h(\rho', \xi(\rho')) \right] \right\} \\ &> P\left\{ \int_s^{\Delta \wedge \rho'} f\,dt + v(\Delta \wedge \rho', \xi(\Delta \wedge \rho')) \right\} \\ &\geqq v'(s, x) \end{aligned}$$

by Proposition 5.9(a). This contradiction implies $\Delta(\zeta) = \rho'(\xi)$ P-a.s. if P is optimal. Consequently, we can work on

$$\Omega' := C \times V$$

and simply delete all reference to S and ζ in the previous discussion. Let us write

$$R''(s, x) = \{P' \in M_1(\Omega') \colon P'(A) = P(A \times Z),\ P \in R_0'(s, x)\}.$$

By the above if $P' \in R''(s, x)$ then it is optimal for the reduced problem; furthermore if

$$S'(\xi, \eta) := \inf\{t \geqq s;\ (t, \xi(t)) \notin D'\}$$

and if P is the induced distribution of $(\xi, \eta, \mathbb{1}_{\{\cdot \geqq S'\}})$ on $\Omega = C \times V \times Z$, then P is an optimal rule for the original problem. Moerover, if P' is Markovian, i.e.,

$$\eta(t) = \delta_{u^*(t, \xi(t))} \quad P'\text{-a.s.}$$

for some Borel function $u^* \colon D' \to U$, then the above equality also holds P-a.s., and

$$\Delta(\zeta) = S'(\xi, \eta) \quad P'\text{-a.s.}$$

Since S' is the first exit time from D', then P is Markovian. Hence it suffices to prove the existence of Markovian optimal controls for the reduced problem.

However the theory of this section requires the set D (in the reduced problem, i.e., D' in the original) to be open in order that a r.c.p.d. P_ω^τ exist, where $\tau = t \wedge y$. Since h is l.s.c., D' is open provided v is u.s.c. This will be the case if

(5.14)
(a) $(s, x, P) \to J(s, P)$ is continuous,
(b) $(s, x) \to R^f(s, x)$ is l.s.c.

In fact in (a) we only require the mapping to be u.s.c., but now (5.3) implies that it is continuous. Recall that a multifunction $x \to A(x)$ is l.s.c. if for any $a_0 \in A(x_0)$ and any sequence $x_n \to x_0$ there exists $a_n \in A(x_n)$ such that $a_n \to a_0$. Conditions such that (5.14)(b) holds are given by El Karoui, Huu Nguyen, and Jeanblanc-Picqué [8, Thm. 5.11(b)] (note that in [8] v is a sup, not an inf). From now on we will simply *assume* (5.14) in the case where $D' \neq D$. Of course if in the original problem stopping is not allowed, i.e., $y = \rho$ P-a.s. for all P, e.g., $h = +\infty$ on D, then $D' = D$ and we do not require (5.14).

We now present a more general version of Proposition 5.9 in the case $\Omega = \Omega' = C \times V$ (i.e., the reduced case). Assume that D' is open and

$$r \colon \bar{D}_A' \to \operatorname{comp}(M_1(\Omega'))$$

is a multifunction such that for all $P \in r(s, x)$

$$P\{\xi(t) = x \colon t < s\} = 1,$$

$$P\{\xi(t) \in A \colon s \leqq t \leqq \rho'(s)\} = 1$$

where $\rho'(s)$ is the first exit time after s of $(t, \xi(t))$ from D'. For $\beta > 0$ and $\tilde{f} \in C_c^+(D')$, the set of nonnegative functions in $C_b(D')$ with compact support, and for $(s, x) \in \bar{D}_A$, $P \in r(s, x)$, define

$$j(s, P) = P \int_s^{\rho'} e^{-\beta t} \tilde{f}(t, \xi(t))\, dt,$$

$$\tilde{v}(s, x) = \inf \{j(s, P): P \in r(s, x)\},$$

$$r'(s, x) = \arg\min \{j(s, P): P \in r(s, x)\}.$$

Then j is l.s.c. on

$$\{(s, x, P): (s, x) \in \bar{D}'_A, P \in r(s, x)\}$$

and $\tilde{v} = 0$ on $\partial D'$. As before we can define τ-transition probabilities Q_ω and measures $\mathring{Q}_\omega^\tau$ and $P/\tau/Q$. Note that if $H \in \text{meas}\,(r)$ then $H(\tau(\omega), \xi(\tau(\omega)))$ is a τ-t.p. and by (5.9) for τ such that $s \leqq \tau \leqq \rho'(s)$ we have

$$Pj(\tau, H(\tau, \xi(\tau))) = j(\tau, P/\tau/H),$$

i.e., the analogue of (5.11) holds. In the case U compact we replace $\bar{D}$ by $\mathbb{R}_+ \times \mathbb{R}^d$ and ρ' by ∞ in the above definition of j. Let us now define

$$\tilde{\Gamma}_t(s, \phi) = \int_s^t e^{-\beta t} \tilde{f}(\theta, \xi(\theta))\, d\theta + \phi(t, \xi(t)), \tilde{\Gamma}_t^*(s, \phi) = \tilde{\Gamma}_{t \wedge \rho'}(s, \phi).$$

We show now that the dynamic programming result (5.12) follows if we assume a variant of the conclusions of Lemmas 5.7 and 5.8 (cf. (5.15)(b), (5.15)(c) below). Then we show using this result that if r satisfies (5.15) then so does r'. Note that we do not require the continuity of ρ'!

PROPOSITION 5.11. *Assume*

(5.15)
- (a) $r: \bar{D}'_A \to \text{comp}\,(M_1(\Omega'))$ *is measurable*;
- (b) *If τ is a finite stopping time such that $s \leqq \tau \leqq \rho'(s)$ and if $P \in r(s, x)$, then there exists a P-null set $N \in \mathcal{A}'_\tau$ such that for $\omega \notin N$, $\mathring{P}_\omega^\tau \in r(\tau(\omega), \xi(\tau(\omega)))$*;
- (c) *if τ is a finite stopping time such that $s \leqq \tau \leqq \rho'(s)$ and if $P \in r(s, x)$, $H \in \text{meas}\,(r)$, then $P/\tau/H \in r(s, x)$.*

Then (5.12) *holds with v replaced by $\tilde{v}$, y by ρ', Γ by $\tilde{\Gamma}$, Γ^* by $\tilde{\Gamma}^*$ and R^f by r.*

Proof. From (5.15)(a) and Lemma 12.1.7 of [19] it follows that $(s, x) \to \tilde{v}(s, x)$ and $(s, x) \to r'(s, x)$ are measurable, hence by Theorem 12.1.10 of [19] there exists a measurable selector H of r' (hence of r). Now the proof goes as in Proposition 5.9 with (5.15)(c) replacing Lemma 5.8 and (5.15)(b) replacing Lemma 5.7. Of course $h = 0$. □

COROLLARY 5.12. *Assume that r satisfies* (5.15). *Then r' satisfies* (5.15).

Proof. Since $r(s, x) \neq \varnothing$, compact and j is l.s.c., then $r'(s, x) \neq \varnothing$, compact. The measurability of r' was established in the proof of Proposition 5.11. Hence r' satisfies (5.15)(a).

Take $P \in r'_0(s, x)$ and σ a finite stopping time, $\tau \leqq \sigma$. Then (5.15)(b) and (5.12)(a) imply that for $\omega \notin N$

$$\tilde{v}(\tau(\omega), \xi(\tau(\omega))) \leqq \mathring{P}_\omega^\tau \tilde{\Gamma}_\sigma(\tau(\omega), \tilde{v}). \tag{5.16}$$

Since $\tilde{\Gamma}_\cdot^*(s, v)$ is a martingale (cf. (5.12)(c)), then

$$\begin{aligned} P\tilde{\Gamma}_\sigma(\tau(\omega), \tilde{v}) &= P\tilde{\Gamma}_\sigma(s, v) \\ &= P\tilde{\Gamma}_\tau(s, \tilde{v}) \\ &= P\tilde{v}(\tau, \xi(\tau)). \end{aligned}$$

This equality implies equality in (5.16) and hence $\mathring{P}^{\tau}_{\omega}\tilde{\Gamma}^{*}_{\cdot}(\tau(\omega), \tilde{v})$ is constant. Assumptions (5.15)(b) and (5.12)(b) now imply that $\tilde{\Gamma}^{*}_{\cdot}(\tau(\omega), \tilde{v})$ is a $\mathring{P}^{\tau}_{\omega}$-martingale, hence by (5.12)(c), $\mathring{P}^{\tau}_{\omega} \in r'_0(\tau(\omega), \xi(\tau(\omega))))$, i.e., (5.15)(b) holds for r'.

Finally, with $P \in r'_0(s, x)$, $s \leqq \tau \leqq \rho'(s)$ and $H \in \operatorname{meas}(r')$, (5.12)(c) implies

$$\begin{aligned}\tilde{v}(s, x) &= P\tilde{\Gamma}_{\tau}(s, \tilde{v}) \\ &= (P/\tau/H)\Gamma_{\tau}(s, \tilde{v}) \\ &= PJ(\tau, H(\tau, \xi(\tau))) \\ &= J(s, P/\tau/H).\end{aligned}$$

Hence $P/\tau/H \in r'(s, x)$. □

Next we show that R'' (or R''_c, i.e., R'_c for the reduced problem) satisfies (5.15).

PROPOSITION 5.13. *Assume* (3.5), (5.2), *and* (5.3). *Then* R'' *satisfies* (5.15).

Proof. Since $\tilde{R}(s, x) \neq \varnothing$ and is compact, by (5.3) and (5.5) $R'_0(s, x)$ is nonempty and compact. The measurability of R'_0 is established in the proof of Lemma 5.5. But $R''(s, x) = \pi(R'_0(s, x))$ where $\pi P(A) := P(A \times Z)$ so π is continuous and hence R'' satisfies (5.15)(a).

We can also show that R'_0 satisfies (5.15)(b) and (5.15)(c). The proof is the same as that of Corollary 5.12 (for r') but using Lemma 5.7 in place of (5.15)(b) (for R_0).

Now for $P' \in R''(s, x)$ set $\pi^{+}P' = P$ where

$$dP(\xi, \eta, \zeta) := \delta_{\zeta'(\xi)}(d\zeta)\, dP'(\xi, \eta)$$

and $\delta_{\zeta'(\xi)}$ is the Dirac measure at $\mathbb{1}_{\{\cdot \geqq \rho'(\xi)\}} := \zeta'(\cdot)$. From our earlier discussion it follows that

$$\pi^{+}: R''(s, x) \to R'_0(s, x),$$

$$\pi\pi^{+}P' = P'.$$

For $P \in R''(s, x)$, let $Q = \pi^{+}P$. By (5.15)(b) for R'_0, for $0 \leqq \tau \leqq \rho'$ there exists a Q-null set $N \in \mathcal{Q}_{\tau}$ such that for $\omega \notin N$

$$\mathring{Q}^{\tau}_{\omega} \in R'_0(\tau(\omega), \xi(\tau(\omega))).$$

Since $\tau \leqq \rho' = \Delta(\zeta)$ Q-a.s. then

$$Q^{\tau}_{\omega} = Q^{\tau}_{(\xi, \eta, \zeta')} \quad Q\text{-a.s.}$$

and it follows that if

$$P^{\tau}_{(\xi, \eta)}(A) := Q^{\tau}_{(\xi, \eta, \zeta')}(A \times Z),$$

then $P^{\tau}_{\omega'} = \pi Q^{\tau}_{(\omega', \zeta')}$ is a r.c.p.d. of P given $\mathcal{Q}'_{\tau}$. Hence for $(\xi, \eta, \zeta') \notin N$

$$\mathring{P}^{\tau}_{(\xi, \eta)} \in R''(\tau(\xi), \xi(\tau(\xi))).$$

Moreover, $(\xi, \eta, \zeta') \in N$ implies that for some P'-null set $N' \in \mathcal{Q}'_{\tau}$,

$$\int_{N_{\omega'}} \delta_{\zeta'}(d\zeta) = 0 \quad \text{if } \omega' \notin N',$$

i.e., $(\omega', \zeta') \notin N$ if $\omega' \notin N'$. Now (5.15)(b) for R'' follows.

Finally, consider (5.15)(c). If $H \in \operatorname{meas}(R'')$, then $\pi^{+}H \in \operatorname{meas}(R'_0)$ since

$$\pi^{+}H(s, x)(A) = H(s, x)(\{(\xi, \eta): (\xi, \eta, \zeta'(\xi)) \in A\}).$$

Now for $Q = \pi^+ P$, $h = \pi^+ H$ and $s \leqq t \leqq \rho'$ we have

$$Q/\tau/h \in R_0'(s, x)$$

by (5.15)(c) for R_0', and hence

$$\pi^+ P/\tau/\pi^+ H \in R_0'(s, x)$$

so the result follows on application of π to this last statement. □

We note that for the case U compact we omit ρ', i.e., we set $\rho' = +\infty$, in the above and replace R_0 by R_c.

Continuing to consider the problem with $\Omega = C \times V$ (i.e., the control runs until time ρ) we now obtain the following proposition.

PROPOSITION 5.14. *Assume* (3.5), (5.2), (5.3), *and* (5.14). *There exists a family of rules* P_{sx}^* *such that* $(s, x) \to P_{sx}^*$ *is measurable,* P_{sx}^* *is in* $R''(s, x)$, *and* $\{P_{sx}^*|_{\mathscr{C}_{\rho'}}\}$ *is a strong Markov process on* $\bar{D}_{A'}$ *where* $P_{sx}^*|_{\mathscr{C}_{\rho'}}$ *is the marginal on* C *restricted to* $\mathscr{C}_{\rho'}$.

Proof. This is almost identical to that of Theorem 12.2.3 of [19]. The idea is to define a sequence of minimization problems: take $\{\beta_n\}$ dense in $(0, \infty)$, $\{f_m\}$ dense in $C_c^+(D')$, and relabel $\{(\beta_n, f_m)\}_{n,m=1}^{\infty} = \{(\beta_l, f_l)\}_{l=1}^{\infty}$. Define inductively

$$R_{l+1}(s, x) = \arg\min \left\{ P \int_s^{\rho'} e^{-\beta_l t} f_l(t, \xi(t))\, dt \right\}$$

with $R_1(s, x) = R''(s, x)$. Then $R_l(s, x)$ is a decreasing sequence and for each l, $R_l(s, x)$ is compact, nonvoid, and R_l satisfies (5.15). The proof is by induction; the case $l = 1$ is given by Proposition 5.13 and the induction step by Corollary 5.12.

If we set

$$R_\infty(s, x) = \bigcap_l R_l(s, x) \subset R''(s, x),$$

then $R_\infty(s, x)$ is nonvoid, compact, and R_∞ satisfies (5.15). Moreover, all elements $R_\infty(s, x)$ have the same marginals on $\mathscr{C}_{\rho'}$. Indeed if P_{sx} and Q_{sx} are in $R_\infty(s, x)$, then for any n, m

$$\begin{aligned} &P_{sx} \int_s^{\rho'} e^{-\beta_n t} f_m(t, \xi(t))\, dt \\ &\quad = Q_{sx} \int_s^{\rho'} e^{-\beta_n t} f_m(t, \xi(t))\, dt \\ &\quad = \min \left\{ \bar{P} \int_s^{\rho'} e^{-\beta_n t} f_m(t, \xi(t))\, dt \colon \bar{P} \in R_l(s, x) \right\} \end{aligned}$$

(if (n, m) correspond to $l+1$). Since $\{\beta_n\}$ is dense in $(0, \infty)$ and since the Laplace transform is unique, then

$$P_{sx}\{[\mathbb{1}_{\{t<\rho'\}} + \mathbb{1}_{\{t\geqq\rho'\}}] f_m(t, \xi(t))\} = Q_{sx}\{[\mathbb{1}_{\{t<\rho'\}} + \mathbb{1}_{\{t\geqq\rho'\}}] f_m(t, \xi(t))\}.$$

If we set

$$\begin{aligned} P_{sx}' &= P_{sx}/\rho'/(\delta_{\xi(\rho')} \times \delta^*), \\ Q_{sx}' &= Q_{sx}/\rho'/(\delta_{\xi(\rho')} \times \delta^*), \end{aligned}$$

then

$$\begin{aligned} P_{sx}' f_m(t, \xi(t)) &= P_{sx} P_{sx}' \{f_m(t, \xi(t)) \,|\, \mathscr{A}_{\rho'}\} \\ &= [P_{sx}\{[\mathbb{1}_{\{t<\rho'\}} + \mathbb{1}_{\{t\geqq\rho'\}}] f_m(t, \xi(t))\} \\ &\qquad + P_{sx}\{[\mathbb{1}_{\{t\geqq\rho'\}} + \mathbb{1}_{\{t>\rho'\}}] f_m(\rho', \xi(\rho'))\}]/2 \\ &= Q_{sx}' f_m(t, \xi(t)) \end{aligned}$$

by the above since $f_m = 0$ on $\partial D'$. It follows that

$$P'_{sx} f(t, \xi(t)) = Q'_{sx} f(t, \xi(t))$$

for all bounded, measurable f, all $(s, x) \in D'_A$ so that (5.15)(b) allows us to extend this equality to finite products of such functions. Hence the finite-dimensional distributions of (ξ, P'_{sx}) and (ξ, Q'_{sx}) are equal and thus $P_{sx} = Q_{sx}$ on $\mathscr{C}_{\rho'} \times \{\varnothing, V\}$.

Finally, if P^*_{sx} is in meas $(R_\infty(s, x))$ then (5.15)(b) and the uniqueness of the marginals on $\mathscr{C}_{\rho'}$ imply that for any stopping time τ, $s \leqq \tau \leqq \rho'(s)$, and $\tilde{A}$ measurable with respect to $\{\xi(\theta)\colon \tau(\omega) \leqq \theta \leqq \rho'(s)\}$

$$(P^*_{sx})^\tau_\omega(\tilde{A}) = P^*_{(\tau(\omega)\xi(\tau(\omega)))}(\tilde{A}) \quad P^*_{sx}\text{-a.s.}$$

Since this is the strong Markov property we are done. □

Note that for the case U compact, we take $f_m \in C^+_c(\mathscr{R}_+ \times \mathscr{R}^d)$. The conclusion of the proposition is the same if we replace ρ' by $+\infty$.

The final step is to construct a feedback control u^* that generates P^*. El Karoui, Huu Nguyen, and Jeanblanc-Picqué [8] appeal to a general theory of Markov processes to do this, but we prefer a more straightforward approach. From (3.10) it follows that for $t \geqq s$

$$\begin{aligned} b(t, \xi(t), \eta(t)) &= \lim_{\theta \downarrow 0} \theta^{-1} P^*_{sx}\{\xi(t+\theta) - \xi(t) \,|\, \mathcal{A}_t\} \quad \text{a.e.,} \\ &= \lim_{\theta \downarrow 0} \theta^{-1} P^*_{t\xi(t)}(\xi(t+\theta) - \xi(t)) \quad P^*_{sx}\, dt\text{-a.s.,} \\ a(t, \xi(t), \eta(t)) &= \lim_{\theta \downarrow 0} \theta^{-1} P^*_{t\xi(t)}\{(\xi(t+\theta) - \xi(t))(\xi(t+\theta) - \xi(t))'\} \quad P^*_{sx}\, dt\text{-a.s.} \end{aligned}$$

Hence if we define on D'_A the Borel measurable functions

$$\begin{aligned} b^*(t, y) &= \lim_{n \to \infty} n P^*_{ty}\left(\xi\left(t + \frac{1}{n}\right) - y\right), \\ a^*(t, y) &= \lim_{n \to \infty} n P^*_{ty}\left\{\left(\xi\left(t + \frac{1}{n}\right) - y\right)\left(\xi\left(t + \frac{1}{n}\right) - y\right)'\right\}, \end{aligned} \tag{5.17}$$

where the functions are defined arbitrarily off their convergence sets, then for $t \geqq s$

$$(a^*(t, \xi(t)), b^*(t, \xi(t))) = (a(t, \xi(t), \eta(t)), b(t, \xi(t), \eta(t))) \quad P^*_{sx}\, dt\text{-a.s.}$$

and a^*, b^* are independent of the initial condition $(s, x) \in D'_A$ since $\tilde{R}(t, y)$ is defined without reference to (s, x).

Similarly, since $\Gamma^t_0(v)$ is a P^*_{0x} martingale then

$$\begin{aligned} f(t, \xi(t), \eta(t)) &= \lim_{\theta \downarrow 0} \theta^{-1} P^*_{sx}\{v((t+\theta), \xi(t+\theta)) - v(t, \xi(t))\} \\ &= \lim_{\theta \downarrow 0} \theta^{-1} P^*_{t\xi(t)}\{v((t+\theta), \xi(t+\theta)) - v(t, \xi(t))\} \\ &= f^*(t, \xi(t)) \quad P^*_{sx}\, dt\text{-a.s.} \end{aligned} \tag{5.18}$$

if we define

$$f^*(t, y) = \lim_{n \to \infty} n P^*_{ty}\left\{v\left(t + \frac{1}{n}, \xi\left(t + \frac{1}{n}\right)\right) - v(t, y)\right\}.$$

THEOREM 5.15. *Assume* (3.4), (3.5), (5.2), (5.3), *and* (5.14). *Then there exists an optimal control in Markov form.*

Note that if the controller cannot stop the process, then (5.14) *is not required* (cf. *Remark* 5.10).

Proof. Let us abbreviate $(a^*, b^*, f^*) = c^*$. From the above it follows that c^* is Borel measurable and if we set

$$N := \{(t, \chi): c^*(t, \chi) \notin K(t, \chi)\}$$

then N is a Borel set. Now by modifying c^* on N we obtain a Borel function $\bar{c}$ such that $\bar{c}(t, \chi) \in K(t, \chi)$ for all $(t, \chi) \in D'_A$. Define

$$N_0 := \{(t, \omega): t \geqq 0, (t, \xi(t)) \in N\}.$$

Then for any (s, x), N_0 has $P^*_{sx}\, dt$ measure zero and for $(t, \omega) \in N_0$, $\bar{c}(t, \xi(t)) = c^*(t, \xi(t))$ (cf. (5.17), (5.18)). From Theorem A.9 applied to $\bar{c}$ with $y = (t, \chi)$ it follows that there exists a Borel function u^* such that for $(s, x) \in D'_A$ and $\phi \in C^2_b(\mathbb{R}^d)$

$$\begin{aligned} &L\phi(t, \xi(t), \eta(t)) = L\phi(t, \xi(t), u^*(t, \xi(t))) \quad P^*_{sx}\, dt\text{-a.e.}, \\ &f(t, \xi(t), \eta(t)) \geqq f(t, \xi(t), u^*(t, \xi(t))) \quad P^*_{sx}\, dt\text{-a.e.} \end{aligned} \tag{5.19}$$

on N_0. Since P^*_{sx} is an optimal rule it follows that u^* is an optimal control process that is feedback or Markovian. Recall that $S = \rho'$ (cf. Remark 5.10). □

Remark 5.16. The following may illuminate the above construction. We write L for the generator when we use c, L^*, when we use c^*, $\bar{L}$, when we use $\bar{c}$, and $L(u^*)$, when we use

$$c(t, \chi, u^*) = (a(t, \chi, u^*), b(t, \chi, u^*), f(t, \chi, u^*)).$$

Then L^* is constructed without reference to (s, x) as are N and N_0, and hence $\bar{c}$ and u^*, i.e., the same control u^* can be used for all initial conditions (s, x). Moreover,

$$\begin{aligned} L(u^*) &= \bar{L} \quad \text{always} \\ &= L^* \quad \text{except on } N_0 \\ &= L \quad \text{by (5.17), (5.18).} \end{aligned}$$

Since N_0 is a $P^*_{sx}\, dt$ null set, then (5.19) holds for any P^*_{sx} in $R_\infty(s, x)$ (cf. the proof of Proposition 5.14 for $R_\infty(s, x)$). Hence u^* is optimal for *any* initial condition (s, x) in D'_A. Moreover it is also optimal for an initial distribution. Indeed if $\mu(\tilde{A})$ is the probability that X_0 is in $\tilde{A}$ with $\{0\} \times \tilde{A} \subset D'_A$, then

$$J(0, P) = \int J(0, P^0_x)\mu(dx)$$

where P^0_x is a r.c.p.d. of P given $X_0 = x$. If R^f_μ are the feasible rules with (3.11) replaced by $P\{\xi(0) \in \tilde{A}\} = \mu(\tilde{A})$, then it follows from (5.6) that

$$\begin{aligned} \inf\{J(0, P): P \in R^f_\mu\} &= \inf\left\{\int J(0, Q(x))\mu\, dx: Q \in \operatorname{meas}(\tilde{R}(0, \cdot))\right\} \\ &= \int v(0, x)\mu(dx) \end{aligned}$$

and since u^* gives the optimal value $v(0, x)$ for *all* x, then it will minimize $J(0, P)$ over R^f_μ.

Finally, we add that if (3.4) fails, then $c^*(t, y)$ lies in $c(t, y, M_1(U))$ almost surely so that we can select a relaxed Markov control $\mu^*(t, y)$ such that off N_s

$$c^*(t) = c(t, \mu^*(t)) \quad P^*_{sx}\, dt\text{-a.e.}$$

Remark 5.17. In the deterministic case, i.e., $a = 0$ Theorem 5.15 says that there exists a Borel measurable function $u^*(t, x)$ such that the marginal on C of the optimal rule P^*_{0x} has support f (call it Σ^*) contained in the set of solutions of

$$\frac{dX}{dt}(t) = b(t, X(t), u^*(t, X(t))) \quad \text{a.e., } X(0) = x. \tag{5.20}$$

No claim is made about uniqueness of such solutions, but Σ^* is nonempty and all elements in Σ^* give rise to the same (minimal) cost. There may of course be solutions of (5.20) that give a larger cost (and do not lie in Σ^*). Hence the conclusion of Theorem 5.15 in the deterministic case is that there exists an optimal pair $(X^*(\cdot), u^*(\cdot, \cdot))$.

Example 5.18. Consider the linear regulator, i.e.,

$$b(t, x, u) = A(t)x + B(t)u, \qquad a(t, x, u) = \sigma(t)\sigma(t)',$$

$$f(t, x, u) = x'M(t)x + u'N(t)u,\ h(t, x) = x'Ex,$$

and $D = (0, T) \times \mathbb{R}^d$, $U = \mathbb{R}^m$. Then $\rho = T$. Assume that A, B, M, N are bounded, measurable, $M(t) \geqq 0$, $N(t) \geqq \alpha I$, $\alpha > 0$, $E \geqq 0$. Then (3.4) holds as does (3.5) with $p < 2$. Note $\gamma = 1$, $\beta = \nu = 0$, so $\bar{p} = p$. It is readily seen that $(x, \omega) \to F(s, x, u)$, i.e., $P \to J_0(s, P)$ is continuous. Since $u \equiv 0$ is feasible then by Theorem 4.7 an optimal control exists. Furthermore, (5.2) is also satisfied with $\lambda(s, x) = K(1 + |x|^2)$ since $u = 0$ is in $\tilde{R}(s, x)$. Moreover, (5.3) holds according to Lemma 5.4 with $u^0 = 0$. Hence there exists an optimal control in Markov form—this is the well-known control $u^*(t, x) = K(t)x$ that can be found by dynamic programming.

However if we now take $a(t, x, u) = \sigma(t, x, u)\sigma(t, x, u)'$ with

$$|\sigma(t, x, u)| \leqq k(1 + |x|^q + |u|^q), \qquad q < 1,$$

and σ Lipschitz in x, then again an optimal Markovian control exists if the convexity hypothesis (3.4) holds. Now $\beta = \nu = 2q < 2$. The case

$$\sigma(t, x, u) = Cx + Du + E,$$

which is *not* covered by this result because $\gamma\bar{\beta} = 2 > p$, is treated by Wonham [20].

It is worth pointing out that if $\rho = T$ and if there exists a constant control process u^0 giving rise to a process $X(t)$ such that (X, u^0) satisfies the hard constraints (if any), and such that $h(t, x) + f(t, x, u^0) \leqq k(1 + |x|^q)$ for some $q < \infty$, then (5.2) holds with $\lambda(s, x) = K(1 + |x|^q)$ as does (5.3) (cf. Lemma 5.4).

Appendix. We will prove some technical lemmas. Let $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$ be an arbitrary probability space with filtration and let $(\Omega, \bar{\mathcal{F}}, \bar{P}, \{\bar{\mathcal{F}}_t\})$ denote its usual augmentation so that it is complete with complete, right-continuous filtration, and $\bar{P}$ is the extension of P to $\bar{\mathcal{F}}$.

LEMMA A.1. *Let R be a measurable, adapted process on $(\Omega, \bar{\mathcal{F}}, \bar{P}, \{\bar{\mathcal{F}}_t\})$ such that for all $T < \infty$*

$$\bar{P} \int_0^T |R_t|\, dt < \infty.$$

Then there exists a predictable process S on $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$ such that $R = S$ except on a (t, ω) set of measure zero.

Proof. Let $\bar{R}$ be a progressively measurable modification of R—it exists since $(\bar{\mathscr{F}}, \bar{\mathscr{F}}_t)$ satisfy the usual hypotheses. Hence for each t

$$\bar{P}\{R_t \neq \bar{R}_t\} = 0$$

and so for all $T < \infty$

$$\begin{aligned}\bar{P}\int_0^T |\bar{R}_t|\,dt &= \int_0^T \bar{P}|\bar{R}_t|\,dt \\ &= \int_0^T \bar{P}|R_t|\,dt \\ &< \infty\end{aligned}$$

by Fubini's Theorem. It follows that there exists a progressive process $\tilde{R}$ indistinguishable from $\bar{R}$ (i.e., $\bar{P}\{\sup_{t\geqq 0}|\tilde{R}_t - \bar{R}_t| = 0\} = 1$) such that for all ω and all $T < \infty$

$$\int_0^T |\tilde{R}_t|\,dt < \infty.$$

Define

$$R'_t = \limsup_{h\downarrow 0} h^{-1}\int_{t-h}^{t} \tilde{R}_\theta\,d\theta.$$

But $\int_{t-h}^{t} \tilde{R}_\theta\,d\theta$ is continuous and adapted, hence predictable, hence so is R'_t. Since $\tilde{R}_\cdot(\omega)$ is locally integrable for each ω, then $R'_t(\omega) = \tilde{R}_t(\omega)$ for t not in some null set N_ω depending on ω, i.e., at all t that are points of approximate continuity of $\tilde{R}_\cdot(\omega)$. Thus R' is a predictable process on $(\Omega, \bar{\mathscr{F}}, \bar{P}, \{\bar{\mathscr{F}}_t\})$ such that R and R' differ only on a (t, ω) null set. According to a result of Dellacherie and Meyer [6, Lemma 7, Appendix 1], there exists a process S, indistinguishable from R', which is predictable on $(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\})$. This is the required process. □

We point out that if R is a.s. continuous, measurable and adapted, then $R' = R$, and so R and S are indistinguishable hence S is continuous almost surely. Thus if R is a Brownian motion on $(\Omega, \bar{\mathscr{F}}, \bar{P}, \{\bar{\mathscr{F}}_t\})$, then S is one on $(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\})$.

The following results are established by Jacod and Mémin [13, Prop. 2.11, Thm. 2.16, Cor. 2.18], and will be used below. We define $dQ = \eta_t(du)\,dt$ and $dQ_n = \eta^n_t(du)\,dt$.

LEMMA A.2. *Assume $Q_n \to Q$ in the stable topology.*

(i) *If $A \subset [0, T]\times U$ is measurable and each t-section of A is closed, then* $\limsup_{n\to\infty} Q_n(A) \leqq Q(A)$.

(ii) *If ϕ is a bounded measurable function defined on $[0, T]\times U$ such that for almost all t the discontinuity set of $\phi(t, \cdot)$ has η_t measure zero, then* $\lim_n \int \phi\,dQ_n = \int \phi\,dQ$.

(iii) *If ψ_n is a sequence of uniformly bounded measurable functions defined on $[0, T]\times U$ such that* $\lim_n Q_n\{(t, u): |\psi_n(t, u)| > \gamma\} = 0$ *for all $\gamma > 0$, then* $\lim_n \int \psi_n\,dQ_n = 0$.

The next result is used in § 4.

LEMMA A.3. *Assume that $(\xi^n, \eta^n) \to (\xi, \eta)$ in $C\times V$ and that $\bar{\phi}$ is a bounded, measurable $\mathbb{R}^k$-valued function defined on $[0, T]\times\mathbb{R}^d\times U$ such that $\bar{\phi}(t, \cdot, \cdot)$ is* l.s.c. *Then*

$$\text{(i)}\quad \lim_n \int_0^T [\bar{\phi}(t, \xi^n_t, \eta^n_t) - \bar{\phi}(t, \xi_t, \eta_t)]^-\,dt = 0$$

where for any function ϕ, $\phi(y)^- := -[\phi(y)\wedge 0]$.

(ii) *If* $\bar{\phi}(t, \cdot, \cdot)$ *is uniformly continuous on* $\mathbb{R}^d \times \{|u| < m\}$ *and if* $\eta_t(\{|u| = m\}) = 0$ *for almost all* $t \in [0, T]$, *then*

$$\lim_{n\to\infty} \int_0^T \int_{|u|<m} \bar{\phi}(t, \xi_t^n, u)\eta_t^n(du)\, dt = \int_0^T \int_{|u|<m} \bar{\phi}(t, \xi_t, u)\eta_t(du)\, dt.$$

Proof. Set $\psi(t, x, u) = \bar{\phi}(t, x, u) - \bar{\phi}(t, \xi_t, u)$ and for $\gamma > 0$

$$A^m = \left\{(t, u) \in [0, T] \times U: \inf_{|y-\xi_t| \leqq 1/m} \psi(t, y, u) \leqq -\gamma\right\}.$$

The l.s.c. of $\bar{\phi}(t, \cdot, \cdot)$ implies that each t-section of A^m is closed. Moreover $A^{m+1} \subset A^m$ and $\cap_m A^m = \varnothing$ since $\bar{\phi}(t, \cdot, u)$ is l.s.c. Now by Lemma A.2(i) we have

$$\lim_m \limsup_n Q_n(A^m) \leqq \lim_m Q(A^m) = 0.$$

Define

$$\begin{aligned} B_n &= \{(t, u): \psi(t, \xi_t^n, u)^- > \gamma\} \\ &= \{(t, u): \psi(t, \xi_t^n, u) < -\gamma\}. \end{aligned}$$

Then $B_n \subset A^m$ for n sufficiently large since $\xi^n \to \xi$. Hence

$$\limsup_n Q_n(B_n) = 0.$$

Now Lemma A.2(iii) implies

$$\lim_n \int_0^T \psi(t, \xi_t^n, \eta_t^n)^-\, dt = \lim_n \int_0^T \int_U \psi(t, \xi_t^n, u)^-\, dQ_n = 0$$

and (i) is established.

Since $\psi(t, \xi_t^n, u)\mathbb{1}_{\{|u|<m\}} \to 0$ uniformly in u for each t in case (ii), then by the bounded convergence theorem

$$\lim_n \int_0^T \int_{|u|<m} \psi(t, \xi_t^n, u)\eta_t^n(du)\, dt = 0$$

and hence

$$\lim_n \int_0^T \int_{|u|<m} \bar{\phi}(t, \xi_t^n, u)\eta_t^n(du)\, dt = \lim_n \int_0^T \int_{|u|<m} \bar{\phi}(t, \xi_t, u)\, dQ^n.$$

Now the result follows from Lemma A.2(ii). This completes the proof. □

Let us next establish the tightness criterion for a set of probability measures on $C \times V \times Z$ used in Proposition 4.5. The result is analogous to one given by Stroock and Varadhan [19, § 1.4], and our proof borrows heavily from theirs.

Writing $\Omega = C \times V \times Z$, we let E be a subset of $M_1(\Omega)$. For μ in $M_1(U)$ recall the notation

$$|\mu|^p = \int_U |\mu|^p \mu(du).$$

Let us set

$$A(T, \delta, \alpha) = \left\{\xi \in C: \sup_{\substack{0 \leqq s \leqq t \leqq T \\ t-s<\delta}} |\xi_t - \xi_s| \leqq \alpha\right\},$$

$$\tilde{A}(T, \delta, \alpha) = A(T, \delta, \alpha) \times V \times Z.$$

LEMMA A.4. *E is tight if for some $p>0$ and all $T<\infty$, $\alpha>0$*

$$\text{(A1)}\qquad \lim_{M\to\infty}\lim_{\delta\downarrow 0}\inf_{P\in E} P\left\{\tilde{A}(T,\delta,\alpha)\cap(\|\xi\|_T\leqq M)\cap\left(\int_0^T|\eta_t|^p\,dt\leqq M\right)\right\}=1,$$

or (if $p=0$)

$$\text{(A1')}\qquad \lim_{M\to\infty}\lim_{\delta\downarrow 0}\inf_{P\in E} P\left\{\tilde{A}(T,\delta,\alpha)(\|\xi\|_T\leqq M)\left(T\int_0^T\eta_t(\{|u|>M\})\,dt\leqq 1\right)\right\}=1.$$

Proof. Fix $\gamma>0$ and let $\alpha=1/n$, $T=n$. Choose M_n, δ_n such that

$$\inf_{P\in E} P\left\{\tilde{A}\left(n,\delta_n,\frac{1}{n}\right)\cap(\|\xi\|_n\leqq M_n)\cap\left(\int_0^n|\eta_t|^p\,dt\leqq M_n\right)\right\}>1-2^{-n}\gamma,$$

which is possible by (A1) if $p>0$. Let

$$A_n=A\left(n,\delta_n,\frac{1}{n}\right)\cap\{\|\xi\|_n\leqq M_n\},$$

$$B_n=\left\{\eta:\int_0^n|\eta_t|^p\,dt\leqq M_n\right\},$$

$$K_n=A_n\times B_n\times Z,\qquad K_\gamma=\bigcap_{n=1}^{\infty}K_n,$$

$$A_\gamma=\bigcap_{n=1}^{\infty}A_n,\qquad B_\gamma=\bigcap_{n=1}^{\infty}B_n.$$

Then $P(K_\gamma)>1-\gamma$ for all P in E. Moreover, by Ascoli's Theorem, A_γ is precompact (i.e., its closure is compact). Recall also that Z is compact.

Since $K_\gamma\subset A_\gamma\times B_\gamma\times Z$ it remains only to show that B_γ, a subset of V, is precompact. But for $t<\infty$ we may set

$$\tilde{\eta}(\cdot)=\int_0^t\eta_\theta(\cdot)\,d\theta$$

and now for η in B_γ

$$\begin{aligned}\tilde{\eta}(\{u\in U:|u|>N\})&\leqq N^{-p}\int_U|u|^p\,d\tilde{\eta}\\&=N^{-p}\int_0^t\int_U|u|^p\eta_\theta(du)\,d\theta\\&=N^{-p}\int_0^t|\eta_\theta|^p\,d\theta\\&\leqq N^{-p}M_n\end{aligned}$$

if $n\geqq t$. Hence $\{\tilde{\eta}:\eta\in B_\gamma\}$ is tight and consequently B_γ is precompact according to the characterization of this property given in § 3.

If $p=0$, we replace $|\eta_t|^p$ by $MT\eta_t(\{|u|>M\})$ in the proof. Only minor modifications are required in the last step. □

Let Λ be a bounded subset of D and suppose

$$E\subset\bigcap_{(s,x)\in\Lambda}R(s,x).$$

We will massage (A1) into a more useful form for such E. Let $q:\mathbb{R}_+\times V\to\bar{\mathbb{R}}_+$ be some progressively measurable function such that for any $\gamma>0$ there exists a finite constant k_γ for which

$$|\eta_t|^p \leqq k_\gamma + \gamma q_t(\eta)$$

for all $\eta\in V$, $t\geqq 0$. For such q and for fixed α, T, M define

$$\sigma_M(\omega)=\inf\left\{t\geqq 0\colon \max\left\{|\xi_t|, \int_0^t |\eta_\theta|^p\,d\theta, \int_0^t q_\theta\,d\theta\right\}\geqq M\right\},$$

$$\xi_t^M=\xi_{t\wedge\sigma_M},\qquad \tau_0(\omega)=0,$$

$$\tau_n(\omega)=\inf\left\{t>\tau_{n-1}(\omega)\colon |\xi_t^M-\xi_{\tau_{n-1}}^M|>\frac{\alpha}{4}\right\},$$

where $\tau_n=+\infty$ if $\tau_{n-1}=+\infty$ or if $|\xi_t^M-\xi_{\tau_{n-1}}^M|<\alpha/4$ for all $t\geqq\tau_{n-1}$. Note that $\tau_n(\omega)$ is strictly increasing to $+\infty$ with n, and if $\tau_n(\omega)<\infty$ then $\tau_n(\omega)\leqq\sigma_M(\omega)$. Now define

$$N(\omega)=\min\{n\colon \tau_{n+1}(\omega)>T\wedge\sigma_M\}=\min\{n\colon \tau_{n+1}(\omega)>T\}$$

$$\delta_\omega(\alpha)=\min\{\tau_n(\omega)-\tau_{n-1}(\omega)\colon 1\leqq n\leqq N(\omega)\},$$

$$\Sigma_M=\{\omega\colon \sigma_M(\omega)\geqq T\},\qquad \Sigma_M^c=\{\omega\colon \sigma_M(\omega)<T\}.$$

In case of (A1′) we take $q_t=0$ and $k_\gamma=1$, and we replace $\int_0^t|\eta_\theta|^p\,d\theta$ by $Mt\int_0^t\eta_\theta(\{|u|>M\})\,d\theta$ in the definition of σ_M and what follows.

LEMMA A.5. *E is tight if for all $T<\infty$, $\alpha>0$ and some $q(\cdot)$ as above*

$$\lim_{M\to\infty}\sup_{P\in E} P\{\Sigma_M^c\}=0 \tag{A2}$$

and for any $M<\infty$

$$\lim_{\delta\downarrow 0}\sup_{P\in E} P\{\delta_\omega(\alpha)\leqq\delta\}=0. \tag{A3}$$

Proof. For ω in Σ_M, $\xi_t^M=\xi_t$ for $t\leqq T$. As is readily seen (cf. [19, Lemma 1.4.1]), for ω in Σ_M

$$\sup_{\substack{0\leqq r\leqq t\leqq T\\ t-r<\delta_\omega(\alpha)}}|\xi_t-\xi_r|=\sup_{\substack{0\leqq r\leqq t\leqq T\\ t-r<\delta_\omega(\alpha)}}|\xi_t^M-\xi_r^M|\leqq\alpha,$$

$$\int_0^T|\eta_t|^p\,dt\leqq M,$$

hence (A1) is implied by

$$\lim_{M\to\infty}\lim_{\delta\downarrow 0}\inf_{P\in E} P\{(\delta_\omega(\alpha)>\delta)\cap\Sigma_M\}=1$$

or by

$$\lim_{M\to\infty}\lim_{\delta\downarrow 0}\sup_{P\in E} P\{(\delta_\omega(\alpha)\leqq\delta)\cap\Sigma_M\}+\lim_{M\to\infty}\sup_{P\in E} P(\Sigma_M^c)=0.$$

The result follows. □

We now make a hypothesis on E, which implies (A3).

(A4) For any $M<\infty$ and any ϕ in $C_b^2(\mathbb{R}^d)$ such that $\phi(x)\geqq 0$, there exists a constant $A_\phi\geqq 0$, which may also depend on M, such that

$$\phi(\xi_t^M)+A_\phi\int_0^{t\wedge\sigma_M}(1+|\eta_\theta|^p)\,d\theta$$

is a nonnegative $(P, \mathcal{A}_t)$ submartingale for every $P \in E$. Moreover, if ψ is a translate of ϕ, then we may take $A_\psi = A_\phi$.

LEMMA A.6. *Assume* (A4). *Then there exists a constant* κ *depending on* M *such that for any* n, γ, δ, $P \in E$

$$P\{\tau_{n+1} - \tau_n \leqq \delta | \mathcal{A}_{\tau_n}\} \leqq \kappa[\delta(1+k_\gamma) + \gamma M]$$

P *almost surely on* $\{\tau_n < \infty\}$.

Proof. Choose ϕ in $C_b^2(\mathbb{R}^d)$ such that $\phi(0) = 1$, $\phi(x) = 0$ for $|x| \geqq \alpha/4$, $0 \leqq \phi(x) \leqq 1$. Let $Q_{\omega'}$ be a r.c.p.d. of P given $\mathcal{A}_{\tau_n}$ and define (n.b. $\omega' = (\xi', \eta', \zeta')$)

$$\phi^{\omega'}(x) = \begin{cases} \phi(x - \xi'_{\tau_n}) & \text{if } \tau_n(\omega') < \infty, \\ 1 & \text{otherwise.} \end{cases}$$

It follows from (A4) (cf. [19, p. 37]) that there exists a P-null set $F \in \mathcal{A}_{\tau_n}$ such that for ω' not in F (and writing σ for σ_M)

$$\phi^{\omega'}(\xi_t^M) + A_\phi \int_0^{t \wedge \sigma} (1 + |\eta_\theta|^p)\, d\theta$$

is a nonnegative $(Q_{\omega'}, \mathcal{A}_t)$ submartingale for $t \geqq \tau_n(\omega')$. Hence for ω' not in F

$$Q_{\omega'}\left\{\phi^{\omega'}(\xi^M_{\tau_{n+1} \wedge (\tau_n(\omega') + \delta)}) + A_\phi \int_0^{\tau_{n+1} \wedge (\tau_n(\omega') + \delta) \wedge \sigma} (1 + |\eta_\theta|^p)\, d\theta\right\}$$

$$\leqq 1 + A_\phi \int_0^{(\tau_n(\omega') \wedge \sigma)} (1 + |\eta_\theta|^p)\, d\theta$$

or indeed

$$\text{(A5)} \quad Q_{\omega'}\{1 - \phi^{\omega'}(\xi^M_{\tau_{n+1} \wedge (\tau_n(\omega') + \delta)})\} \leqq A_\phi \left[\delta + Q_{\omega'}\left\{\int_{\tau_n(\omega') \wedge \sigma}^{\tau_{n-1} \wedge (\tau_n(\omega') + \delta) \wedge \sigma} |\eta_\theta|^p\, d\theta\right\}\right].$$

But if $\tau_n(\omega') < \infty$ then $\tau_{n+1} \leqq \tau_n + \delta$ only if $\tau_{n+1} \leqq \sigma$ and $1 - \phi^{\omega'}(\xi^M_{\tau_{n+1}} \wedge (\tau_n(\omega') + \delta)) = 1$, so that the left side of (A5) is greater than $Q_{\omega'}\{\tau_{n+1} - \tau_n \leqq \delta\}$. Moreover, under the integral on the right $\theta \leqq \sigma$, hence the integral is bounded by $\delta k_\gamma + \gamma M$. The result follows with $\kappa = A_\phi$. □

LEMMA A7. *Assume* (A4). *Then* (A3) *holds.*

Proof.

$$P\{\delta_\omega(\alpha) \leqq \delta\} \leqq P\left\{\min_{1 \leqq i \leqq k} \tau_i - \tau_{i-1} \leqq \delta\right\} + P\{N > k\}$$

$$\leqq \sum_{i=1}^{k} P\{\tau_i - \tau_{i-1} \leqq \delta\} + P\{N > k\}$$

$$\leqq k\kappa[\delta(1+k_\gamma) + \gamma M] + P\{N > k\}$$

by Lemma A.6. Since we can take $\gamma \downarrow 0$ we need only show that

$$\text{(A6)} \qquad \limsup_{k \to \infty} \sup_{P \in E} P\{N > k\} = 0.$$

But by Lemma A.6

$$P\{e^{-(\tau_{i+1} - \tau_i)} | \mathcal{A}_{\tau_i}\} \leqq P\{\tau_{i+1} - \tau_i \leqq r | \mathcal{A}_{\tau_i}\} + e^{-r} P\{\tau_{i+1} - \tau_i > r | \mathcal{A}_{\tau_i}\}$$

$$\leqq e^{-r} + (1 - e^{-r}) P\{\tau_{i+1} - \tau_i \leqq r | \mathcal{A}_{\tau_i}\}$$

$$\leqq e^{-r} + (1 - e^{-r})\kappa[r(1+k_\gamma) + \gamma M] \quad P \text{ a.s.}$$

$$:= \tilde{\lambda} < 1$$

for suitable γ and $r>0$. Now Lemma 1.4.5 of [19] implies that

$$P\{N \geqq k\} \leqq e^{\gamma} \tilde{\lambda}^k,$$

which guarantees (A6). □

Combining Lemmas A.4 and A.6 gives Theorem A.8.

THEOREM A.8. *Let E be a subset of $\bigcup_{(s,x)\in\Lambda} R(s, x)$ that satisfies* (A2) *and* (A4). *Then E is tight.*

We turn now to proving a measurable selection theorem that was used in Theorems 3.6 and 5.15. U remains as before a closed subset of a Euclidean space. Let $(Y, \mathcal{F}, P)$ be a measure space ($\mathbb{R}_+ \times \Omega$ with the progressively measurable σ-algebra in Theorem 3.6 and D_A with the Borel σ-algebra in Theorem 5.15) and for some natural numbers k, m let

$$c^1: Y \to \mathbb{R}^k, \quad c^2: Y \to \mathbb{R}^m, \quad \phi: Y \times U \to \mathbb{R}^k, \quad \psi: Y \times U \to \mathbb{R}^m_+$$

be given measurable functions with $u \to \phi(y, u)$ continuous and $u \to \psi_i(y, u)$ l.s.c. for each y and each i in $\{1, 2, \cdots, m\}$. Define

$$K(y) = \{(\phi(y, u), z) \in \mathbb{R}^k \times \mathbb{R}^m : z_i \geqq \psi_i(y, u), u \in U\}.$$

THEOREM A.9. *If $(c^1(y), c^2(y))$ lies in $K(y)$ for each y, then there exists a measurable function $u: Y \to U$ such that*

$$c^1(y) = \psi(y, u(y)),$$
$$c_i^2(y) \geqq \phi_i(y, u(y)), \qquad i = 1, \cdots, m.$$

Proof. Define

$$A(y) = \{u \in U : c^1(y) = \phi(y, u), c^2(y) \geqq \psi(y, u)\}$$

where $a \geqq b$ means $a_i \geqq b_i$ for all i if $a, b \in \mathbb{R}^m$. We must show that A has a measurable selector. Let $U_N = \{u \in U : |u| \leqq N\}$ and let

$$A_N(y) = \{u \in U_N : c^1(y) = \phi(y, u), c^2(y) \geqq \psi(y, u)\}.$$

Then U_N is compact, and for each y, $A_N(y)$ increases to $A(y)$. Moreover, the hypothesis of the theorem implies that $A(y) \neq \varnothing$ so if

$$B_N = \{y : A_N(y) \neq \varnothing\}, \qquad N = 1, 2, \cdots,$$

then the B_N are measurable sets increasing to Y.

Suppose A_N restricted to B_N has a measurable selector u_N, $N \geqq 1$. With $B_0 = \varnothing$, define

$$u(y) = \sum_{N=1}^{\infty} u_N(y) \mathbb{1}_{B_N \setminus B_{N-1}}(y).$$

Then u is a measurable selector of $A(\cdot)$ and hence without loss of generality *we may take U compact.*

We apply a result of Dynkin and Yushkevich [21, pp. 57–58], according to which A admits a measurable selection if

(a) $A(y)$ is nonempty and compact for each y,

(b) there exists a sequence of open sets

$$Q^1(y) \supset Q^2(y) \supset \cdots \supset A(y)$$

such that

(i) For any n, z the set $\{y : z \in Q^n(y)\}$ is measurable,

(ii) Every sequence $\{z^n\}$ with $z^n \in Q^n(y)$, has a limit point in $A(y)$.

As noted above $A(y)$ is a nonempty subset of the compact set U so (a) holds if $A(y)$ is closed. But if $z_n \to z_0$ with z_n in $A(y)$, then by continuity of ϕ and l.s.c. of ψ_i

$$\phi(y, z_0) = \lim_n \phi(y, z_n) = c^1(y),$$

$$\psi_i(y, z_0) \leqq \liminf_n \psi_i(y, z_n) \leqq c_i^2(y), \qquad i = 1, 2, \cdots, m,$$

so that $z_0 \in A(y)$, i.e., (a) holds.

Since ψ_i is measurable and l.s.c. in u, there exist $\mathbb{R}_+$-valued, measurable functions ψ_i^n, continuous in U, such that ψ_i^n increases to ψ_i pointwise (cf. [21, p. 51]). Define

$$Q^n(y) = \{u \in U: |c^1(y) - \phi(y, u)| < 1/n,\ c_i^2(y) > \psi_i^n(y, u) - 1/n,\ i = 1, \cdots, m\}.$$

This gives a decreasing sequence of open sets containing $A(y)$. Moreover, the measurability of

$$\{y: z \in Q^n(y)\}$$

follows from that of c^1, c^2, ϕ, and ψ^n. Finally, if $z^n \in Q^n(y)$ then by compactness of U there exists a subsequence (again denoted by $\{z^n\}$) converging to z in U. We want to show that $z \in A(y)$. From the continuity of ϕ and ψ_i^n it follows that

$$|c^1(y) - \phi(y, z)| = \lim_n |c^1(y) - \phi(y, z^n)| \leqq \lim_n \frac{1}{n} = 0,$$

$$\psi_i^m(y, z) = \lim_n \left[\psi_i^m(y, z^n) - \frac{1}{n}\right].$$

But ψ_i^m increases with m so for $n > m$

$$\psi_i^m(y, z^n) - \frac{1}{n} \leqq \psi_i^n(y, z^n) - \frac{1}{n} < c_i^2(y)$$

and hence $\psi_i^m(y, z) \leqq c_i^2(y)$. Finally,

$$\psi_i(y, z) = \lim_m \psi_i^m(y, z) \leqq c_i^2(y)$$

and hence $z \in A(y)$, i.e., (b) holds. The result follows. □

LEMMA A.10. *Assume that $Q: \bar{D} \to M_1(\Omega)$ is measurable and let $Q'(s, x)$ be a* r.c.p.d. *of $Q(s, x)$ given $\zeta(s) = 1$. Then Q' is measurable.*

Proof. Let K be any closed set in Ω. It suffices to show that $(s, x) \to Q'(s, x)(K)$ is measurable (cf. [2]) Appendix III. Let $\{\pi^k\}$ be a sequence of partitions of $\bar{\mathbb{R}}_+$ such that diam $(\pi^k) \to 0$ and π^{k+1} is a refinement of π^k. If $\pi^k = \{s_n^k\}$, then

$$\begin{aligned} Q(s, x)(K \cap \{\zeta(s) = 1\} &= Q(s, x)(K \cap \{\Delta(\zeta) \leqq s\}) \\ &= \lim_k \sum_n \mathbb{1}_{\{s_m^k \leqq s\}} [Q(s, x)(K \cap \{s_n^k \leqq \Delta(\zeta)\}) \\ &\quad - Q(s, x)(K \cap \{s_{n+1}^k \leqq \Delta(\zeta)\})] \end{aligned}$$

and hence

$$(s, x) \to Q(s, x)(K \cap \{\zeta(s) = 1\}) = Q(s, x)(\{\zeta(s) = 1\}) Q'(s, x)(K)$$

is measurable since Q is. Moreover, $(s, x) \to Q(s, x)(\{\zeta(s) = 1\})$ is also measurable, hence so is $Q'(s, x)(K)$. □

LEMMA A.11. *Let $s \geqq 0$ and suppose that P is a probability measure on $(\Omega, \mathcal{A}^s)$. If $\omega' \in \Omega$ and*

$$P\{\omega: (\xi(s), \zeta(s)) = (\xi'(s), \zeta'(s))\} = 1,$$

then there exists a unique probability measure $\delta_{\omega'}/s/P$ on $(\Omega, \mathcal{A})$ such that

$$(\delta_{\omega'}/s/P)\{\omega: \omega(t) = \omega'(t),\ t \leqq s\} = 1,$$

$$(\delta_{\omega'}/s/P)(A) = P(A) \quad \text{if } A \in \mathcal{A}^s.$$

Proof. The proof goes much like that of Lemma 6.1.1 of [19] so we only sketch it. If I is a subinterval of $[0, \infty)$, write $V(I)$ for the measurable functions $I \to M_1(U)$ with topology induced by i_I where

$$i_I(\eta)(\cdot) = \int_{In[0,\cdot]} \eta(\theta)\, d\theta \in C(\mathbb{R}_+; M_1(U))$$

(cf. § 3.10). Similarly, write $Z(I)$ for the set of right-continuous nondecreasing functions: $I \to \{0, 1\}$ with topology induced by Δ_I where

$$\Delta_I(\zeta) = \inf\{t \in I: \zeta(t) = 1\} \in \bar{\mathbb{R}}_+$$

(cf. § 3.11). Now define

$$\Phi: \Omega \to C([s, \infty); \mathbb{R}^d) \times i_{[s,\infty)} V([s, \infty)) \times Z([s, \infty)) := Y_\infty$$

by

$$\Phi(\omega)(t) = \left(\xi(t), \int_s^t \eta(\theta)\, d\theta, \zeta(t)\right), \qquad t \geqq s,$$

and define

$$Y_0 = C([0, s]; \mathbb{R}^d) \times i_{[0,s]} V([0, s]) \times Z([0, s]),$$

$$\tilde{X} = Y_0 \times Y_\infty,$$

$$X = \{(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3) \in \tilde{X}: \alpha_1(s) = \beta_1(s),\ \alpha_3(s) = \beta_3(s)\},$$

$$\Psi(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3)(t) = \begin{cases} (\alpha_1(t), \alpha_2(t), \alpha_3(t)), & t \leqq s, \\ (\beta_1(t), \alpha_2(s) + \beta_2(t), \beta_3(t)), & t \geqq s. \end{cases}$$

The proof is now identical to that of Lemma 6.1.1 of [19], i.e., if $\tilde{\delta}_{\omega'}$ is the point mass on Y_0 at $\tilde{\omega}$ where $\tilde{\omega}(t) = \tilde{\omega}'(t)$, $0 \leqq t \leqq s$, if $\tilde{P} = \tilde{\delta}_{\omega'} \times P \circ \Phi^{-1}$ on $\tilde{X}$ then $\tilde{P}(X) = 1$ and $\tilde{P} \circ \Psi^{-1}$ is the desired measure $\delta_{\omega'}/s/P$. □

Let us finally consider the continuity of ρ following an idea of Lions and Menaldi [22]. We assume that $\partial D \neq \varnothing$ (otherwise $\rho = +\infty$), and we define

$$D_N = \{(t, x) \in D: |(t, x)| < N\}.$$

We say that D has the *local uniform exterior sphere property* if

$$\forall N, \quad \forall \xi \in \partial D \cap \bar{D}_N \quad \text{there exist } r_N, \xi' \notin \bar{D}$$

such that

$$\{y: |y - \xi'| \leqq r_N\} \cap \bar{D} = \{\xi\}.$$

Hence D_N has the uniform exterior sphere property for every $N > 0$. Note that $\bar{n}(\xi) := (\xi' - \xi)/r_N$ defines a unit outward normal at ξ.

Let $\alpha = (\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{X_t\}, \{\mu_t\}, S)$ be a relaxed control. Since we are only interested in the continuity of ρ on $\{\rho = S\}$ we assume $\rho \leqq S$. For notational convenience we will consider the autonomous case (this can always be arranged since σ may be nonsingular). Let τ_x stand for the first exit time from $\bar{D}$ and τ_x^N for that from $\bar{D}^N$ for the process $\bar{X}$ defined below with $\bar{X}(0) = x$. Note that $\bar{D}$ denotes the closure of D.

THEOREM A.12. *Assume*

(i) *D has the local uniform exterior sphere property.*

(ii) *For all N, there exists $\bar{\mu}_N(x): \bar{D}_N \to M_1(U)$ such that*

$$d\bar{X}_t = b(\bar{X}_t, \bar{\mu}_N(\bar{X}_t))\, dt + \sigma(\bar{X}_t, \bar{\mu}_N(\bar{X}_t))\, dw_t, \qquad \bar{X}_0 = x$$

has, for each $x \in \bar{D}_N$, a Markov solution on $0 \leqq t \leqq \tau_x^N$. Write $\bar{b}_N(x)$ for $b(x, \bar{\mu}_N(x))$ and similarly for $\bar{\sigma}_N(x)$.

(iii) *$\bar{a}_N$ and $\bar{b}_N$ are uniformly continuous in a neighbourhood of $\partial \bar{D}^N \cap \partial D$ (each N) and are bounded on $\bar{D}_N$.*

(iv) *$\partial D = \Gamma_0 \cup \Gamma_1 \cup \Gamma_2$ where*

$$P\{X_S \in \Gamma_0\} = 0$$

and for all N there exists $\alpha_N > 0$ such that for all $x \in \Gamma_1 \cap \bar{D}_N$

$$\operatorname{tr} \bar{a}_N(x) - 2r_N \bar{n}(x) \cdot \bar{b}_N(x) \leqq -\alpha_N,$$

and for all $x \in \Gamma_2 \cap \bar{D}_N$

$$\bar{n}(x) \cdot \bar{a}_N(x) \bar{n}(x) \geqq \alpha_N.$$

Then ρ is continuous.

Proof. We will first show that if $\bar{X}(0) \in \Gamma_1 \cup \Gamma_2$, $|\bar{X}(0)| \leqq N$, then $\bar{X}$ leaves $\bar{D}_N$ immediately. For $\xi \in (\Gamma_1 \cup \Gamma_2) \cap \bar{D}_N$, $x \in \bar{D}_N$ define a function

$$W(x, \xi) := \exp(-kr_N^2) - \exp(-k|x - \xi'|^2)$$

with $k > 0$ to be chosen later and ξ', r_N defined by the external sphere property. This property implies $W(x, \xi) \geqq 0$ with equality if and only if $x = \xi$. Then $x \to W(x, \xi)$ is in $C_b^2(\bar{D}_N)$ and $1 \geqq W(x, \xi)$. Write

$$\| W(\cdot, \xi) \|_2 = \sup_{x \in \bar{D}_N} \{ W(x, \xi) + |W_x(x, \xi)| + |W_{xx}(x, \xi)| \}.$$

Moreover, given $\varepsilon > 0$ there exists $\alpha' > 0$ such that if $|x - \xi| \geqq \varepsilon$ then $W(x, y) \geqq \alpha'$.

We now claim that given N we can choose $k < \infty$, $\alpha'' > 0$, $\varepsilon > 0$, such that $\mathcal{L}W(x, \xi) \leqq -\alpha''$ if $|x - \xi| < \varepsilon$ where $\mathcal{L}$ is the generator of $\bar{X}$. Observe first that

$$\mathcal{L}W(x, \xi) = k\, e^{-k|x-\xi'|^2} [\operatorname{tr} \bar{a}_N(x) + 2(x - \xi') \cdot \bar{b}_N(x) - 2k(x - \xi') \cdot \bar{a}_N(x)(x - \xi')].$$

If $\xi \in \Gamma_1$ then

$$\begin{aligned}
\mathcal{L}W(x, \xi) &\leqq k\, e^{-k|x-\xi'|^2} [\operatorname{tr} \bar{a}_N(x) + 2(x - \xi') \cdot \bar{b}_N(x)] \\
&= k\, e^{-k|x-\xi'|^2} [\operatorname{tr} \bar{a}_N(\xi) + 2(\xi - \xi') \cdot \bar{b}_N(\xi) + o(1)] \\
&\leqq k\, e^{-k|x-\xi'|^2} [-\alpha_N + o(1)] \\
&\leqq k\, e^{-k|x-\xi'|^2} (-\alpha_N / 2) \\
&\leqq -\frac{1}{2} \alpha_N k\, e^{-k(r_N + \varepsilon)^2}
\end{aligned}$$

for ε sufficiently small.

If $\xi \in \Gamma_2$ then

$$\begin{aligned}\mathscr{L}W(x,\xi) &\leqq k\, e^{-k|x-\xi'|^2}[-2k(\xi-\xi')\cdot \bar{a}_N(\xi)(\xi-\xi')+2ko(1)+K_N]\\ &\leqq k\, e^{-k|x-\xi'|^2}[-kr_N^2\alpha_N+K_N]\\ &\leqq k\, e^{-k|x-\xi'|^2}(-\alpha_N/2)\end{aligned}$$

if ε is sufficiently small and k sufficiently large. The result follows.

We can now conclude that with k chosen as in the claim, there exists $c>0$ such that

$$\mathscr{L}W(x,\xi)-cW(x,\xi)\leqq -\alpha''.$$

For $|x-\xi|<\varepsilon$ this follows from the claim, and for $|x-\xi|\geqq\varepsilon$, α' given by this ε (cf. above)

$$\begin{aligned}\mathscr{L}W(x,\xi)-cW(x,\xi) &\leqq O(\|W(\cdot,\xi)\|_2)-c\alpha'\\ &\leqq -\alpha''\end{aligned}$$

if c is sufficiently large.

Now consider $\bar{X}$ with $\bar{X}(0)=\xi$. Then

$$d(W(\bar{X}(t),\xi)\,e^{-ct})=[\mathscr{L}W(\bar{X}(t),\xi)-cW(\bar{X}(t),\xi)]\,e^{-ct}\,dt+O(1)\,dw_t,$$

$$\frac{d}{dt}EW(\bar{X}(t),\xi)\,e^{-ct}\leqq -\alpha''$$

so that

$$0\leqq EW(\bar{X}(\tau^N),\xi)\,e^{-c\tau^N}\leqq -\alpha''(1-E\,e^{-c\tau^N})/c,$$

i.e., $E\,e^{-c\tau^N}\geqq 1$ and hence $\tau^N=0$ almost surely.

Let us finally consider the process $\{X_t\}$ corresponding to the relaxed control α. Fix N and let ρ^N denote the first exit time of X from D_N. On the set $\{\rho=\rho^N<\infty\}$ we have $X_\rho\in\Gamma_1\cup\Gamma_2\subset\partial D$ and we can consider the process $\bar{X}$ with $\bar{X}(0)=X_\rho$. Define

$$\tilde{X}_t=\begin{cases}X_t, & t\leqq\rho,\\ \bar{X}_{t-\rho}, & t\geqq\rho.\end{cases}$$

If $\tilde{\tau}^N$ is the first exit time of $\tilde{X}$ from $\bar{D}_N$ and $\tau^N_{X(\rho)}$ that of $\bar{X}$, then on $\{\rho=\rho^N<\infty\}$

$$\tilde{\tau}^N=\rho^N+\tau^N_{X(\rho)}=\rho^N \quad \text{a.s.}$$

where the last equality follows from our previous labours. Since $\tilde{\tau}^N$ is u.s.c. and ρ^N is l.s.c. then ρ is almost surely continuous on $\{\rho=\rho^N<\infty\}$. But $\{\rho^N=\rho<\infty\}\uparrow\{\rho<\infty\}$ since for $\omega\in\{\rho<\infty\}$ if $N>|X(\rho)|$ then $\rho^N=\rho$. Hence ρ is almost surely continuous on $\{\rho<\infty\}$ and we are done. □

REFERENCES

[1] V. E. BENES, *Existence of optimal stochastic controls*, SIAM J. Control, 9 (1971), pp. 446–472.

[2] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[3] J. M. BISMUT, *Théorie probabiliste du contrôle des diffusions*, Mem. Amer. Math. Soc., 4 (1976).

[4] V. S. BORKAR, *The probabilistic structure of controlled diffusion processes*, preprint.

[5] M. H. A. DAVIS, *On the existence of optimal policies in stochastic control*, SIAM J. Control, 11 (1973), pp. 587–594.

[6] C. DELLACHERIE AND P. A. MEYER, *Probabilités et potentiel*, Vols. I, II, Hermann, Paris, 1975, 1980.

[7] N. EL KAROUI, *Les aspects probabiliste du contröl stochastique*, Ecole d'été de Saint-Flour, 1979, Lecture Notes in Mathematics 876, Springer-Verlag, Berlin, New York, 1981, pp. 74–239.

[8] N. EL KAROUI, D. HUU NGUYEN, AND M. JEANBLANC-PICQUÉ, *Compactification methods in the control of degenerate diffusions: existence of an optimal control*, Stochastics, 20 (1987), pp. 169-219.

[9] W. H. FLEMING AND M. NISIO, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777-794.

[10] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[11] U. G. HAUSSMANN, *Existence of optimal Markovian controls for degenerate diffusions*, Lecture Notes in Control and Information Sciences 78, Springer-Verlag, Berlin, New York, 1986, pp. 171-186.

[12] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.

[13] J. JACOD AND J. MÉMIN, *Sur un type de convergence intermédiare entre la convergence en loi et la convergence en probabilité*, Séminaire de Probabilité 15, Lecture Notes in Mathematics 850, Springer-Verlag, Berlin, New York, 1981, pp. 529-546.

[14] N. V. KRYLOV, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.

[15] H. J. KUSHNER, *Existence results for optimal stochastic controls*, J. Optim. Theory Appl., 15 (1975), pp. 347-359.

[16] P. LOEWEN, *Existence theory for a stochastic Bolza problem*, I.M.A. J. Math. Control Inform. 4 (1987), pp. 301-320.

[17] M. MÉTIVIER, *Semimartingales*, de Gruyter, Berlin, 1982.

[18] M. SION, *A Theory of Semigroup Valued Measures*, Lecture Notes in Mathematics 355, Springer-Verlag, Berlin, New York, 1973.

[19] D. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.

[20] W. M. WONHAM, *Random differential equations in control theory*, in Probabilistic Methods in Applied Mathematics, Vol. II, A. T. Bharucha-Reid, ed., Academic Press, New York, 1969.

[21] E. B. DYNKIN AND A. A. YUSHKEVICH, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.

[22] P. L. LIONS AND J. L. MENALDI, *Optimal control of stochastic integrals and Hamilton-Jacobi-Bellman equations*, SIAM J. Comput., 20 (1982), pp. 58-95.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 903-924, July 1990

007

GRADED APPROXIMATIONS AND CONTROLLABILITY ALONG A TRAJECTORY*

ROSA MARIA BIANCHINI† AND GIANNA STEFANI‡

Abstract. Graded approximations of an affine control system are defined and their properties are investigated. In particular, it is proved that the local controllability along a reference trajectory of an approximating system implies the local controllability along the corresponding reference trajectory of the original system.

Using graded approximations, sufficient conditions of local controllability along a reference trajectory that generalize some known results are given.

Key words. affine control systems, graded structure, graded approximations, local controllability

AMS(MOS) subject classifications. 93B05, 93C10, 93C45

Introduction. In this paper we give sufficient conditions for a reference trajectory $t \mapsto \hat{x}(t)$ of a control system to be in the interior of the reachable set at each time t. If this is the case, we say that the control system is locally controllable along $\hat{x}$. This property is related to optimal control problems by the fact that an optimal trajectory usually belongs to the boundary of the reachable sets. Therefore necessary conditions for $\hat{x}$ to be optimal can be derived from sufficient conditions of local controllability. Moreover, the study from a variational point of view of the local controllability property may provide tools for obtaining high-order variations and hence high-order maximum principles. In this context it seems more interesting to also consider the cases in which the reachable sets have empty interiors, but it makes sense to define a "relative interior."

If $\hat{x}(t)$ belongs to this relative interior at each time, we say that the control system is weakly locally controllable along $\hat{x}$. Note that this weaker property allows us to reduce the study of local controllability along a trajectory relative to a C^∞ control map to the study of local controllability of a new system along a trajectory relative to a constant control. In fact, this can be done adding one dimension to the state space, but the reachable sets of this new system always have empty interiors. In a forthcoming paper, the authors will show how the results on weak local controllability can be used to obtain high-order variations for a suitable high-order maximum principle. Some of these results have been announced at the 8th International Symposium on the Mathematics of Networks and Systems, held in Phoenix in June 1987 [23].

If $\hat{x}$ is stationary, i.e., $\hat{x}(t) \equiv \xi_0$, the property under consideration reduces to small time local controllability at ξ_0. This last property is related to the minimum time problem and to the global controllability property. We will study the local controllability property in the framework of the geometric theory in line with the philosophy that the local geometric properties of a control system must be described by the "relations at the initial point" between the vector fields belonging to the Lie algebra associated to the system itself. This point of view has provided several important results on this topic (see, for example, [24], [18], [6], etc. for the stationary case, and [1], [10], [11], [28], etc. for the general case). Most of the results point out some elements of the Lie algebra associated with the system as "possible obstructions" and give conditions to neutralize them. Nevertheless in our opinion the local controllability property is not yet completely understood.

* Received by the editors November 16, 1988; accepted for publication (in revised form) June 29, 1989.
† Istituto di Matematica U. Dini, Viale Morgagni 67/a, 50134 Firenze, Italy.
‡ Dipartimento di Matematica e Applicazioni, Via Mezzocannone 8, 80134 Napoli, Italy.

To get a deeper analysis of the property, we consider a class of very simple systems for which it is easy to understand the reason only some particular brackets can be obstructions. By this we mean that if these brackets vanish at the initial point then the system is locally controllable along the reference trajectory.

We study separately a class of perturbations that do not destroy the property. Note that most of the known sufficient conditions guarantee that the property is preserved under suitable perturbations of the system, but it is not clear in general in what sense the property may be considered "stable" [3], [17]. For example, the small time local controllability of the system obtained by taking a Taylor approximation of the vector fields does not guarantee the local controllability of the original system (see Example 3.1.).

Finally, we will look for conditions under which a system can be considered an admissible perturbation of a "simple" one for which the obstructions vanish.

The perturbations under which the local controllability property is stable are defined in the framework of graded structures. Graded approximations of vector fields and control systems have been considered by several authors (see, for example, [4], [7], [9], [14], [19]–[21], etc.). It is the opinion of the authors that the graded approximations of control systems are interesting in themselves and that they can also be usefully applied in other control problems. This point of view allows us to unify and to generalize several results on local controllability in the stationary and the nonstationary case both with bounded and unbounded controls.

The plan of the paper is the following. In § 1 we revisit some known results about local controllability for the class of analytic nilpotent systems from the point of view described above. We define the obstructions and state our main result (Theorem 1.1) on weak local controllability along a reference trajectory. Finally, we give some examples and comparisons with known results.

In § 2 we recall some definitions and properties from the theory of the graded structures and we prove the main approximation result (Theorem 2.2). Theorem 2.2 is crucial in the proof of the main result and it will be used to give a variational version of Theorem 1.1 in a forthcoming paper. It is the opinion of the authors that the approximation result is interesting by itself. In fact, it can be thought of as a sort of generalization of the linearization principle for the stationary case (see Corollary 2.3). Note that in Theorem 2.2 there are no assumptions on the constraint set Ω; hence it can be used to test the local controllability property also in the case of one-sided controls.

In § 3 the graded approximation induced by a filtration of a Lie algebra is studied. In particular, we prove that the Lie algebra associated with the graded approximating system is nilpotent. In § 4 we prove Theorem 1.1. In the Appendix we recall some properties of the analytic systems. We also prove a result (Lemma A) that is probably known but that the authors did not find in the literature in an appropriate version.

Let us remark that the results in §§ 2 and 3 are similar to the ones stated by Stefani in [21]. The main difference is due to the fact that the trajectory may be not stationary.

1. Notation and statement of the main result. In the sequel we will use the following notation:

M is a C^∞, finite-dimensional, paracompact, connected manifold.

$\mathcal{F}(M)$ is the algebra of C^∞ real functions on M.

$\mathcal{V}(M)$ is the Lie algebra of C^∞ vector fields on M.

If $f \in \mathcal{V}(M)$, $(t, \xi) \mapsto \exp tf \cdot \xi$ denotes the local flow of f and $\mathrm{ad}_f : \mathcal{V}(M) \to \mathcal{V}(M)$, $g \mapsto \mathrm{ad}_f g \equiv [f, g]$ denotes the Lie derivation in $\mathcal{V}(M)$ with respect to f.

$\mathcal{D}(M)$ is the noncommutative algebra of the C^∞ differential operators on $\mathcal{F}(M)$.

If $S, T \in \mathcal{D}(M)$ and $\alpha \in \mathcal{F}(M)$, $S \cdot T$ denotes the composition between S and T and $S \cdot \alpha$ the function obtained applying S to α.

It is known that each $f \in \mathcal{V}(M)$ may be considered as an element of $\mathcal{D}(M)$. If $x = (x^1, \cdots, x^n)$ is a chart at a point ξ_0, we will write f in coordinates as $f = \sum_{i=1}^n f^i(\partial/\partial x^i)$, so that if $\alpha \in \mathcal{F}(M)$, $f \cdot \alpha = \sum_{i=1}^n f^i(\partial\alpha/\partial x^i)$.

If F is a subset of $\mathcal{V}(M)$, Lie F will denote the Lie subalgebra of $\mathcal{V}(M)$ generated by F, and $\mathcal{D}(F)$ will denote the subalgebra of $\mathcal{D}(M)$ generated by F.

Let $h: M \to M$ be a diffeomorphism; $h_*: TM \to TM$ is the tangent map of h.

Let Ω be a given subset of $\mathbb{R}^m$ such that $0 \in \Omega$ and span $\Omega = \mathbb{R}^m$. To each family $\mathbf{f} = (f_0, f_1, \cdots, f_m)$ of C^∞ vector fields on M, we associate the affine control process (Σ_f, Ω) on M, where the state x satisfies the equation

$$(\Sigma_{\mathbf{f}}) \qquad \dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$$

and the control u belongs to the class $\mathcal{U}$ of piecewise constant maps from $\mathbb{R}$ into Ω.

Let $\xi \in M$; we denote by $s_{\mathbf{f}}(t, \xi, u)$ the value at time t of the solution of $(\Sigma_{\mathbf{f}})$ relative to the control u, starting at ξ and we denote by $R_{\mathbf{f}}(\xi, t)$ the reachable set from ξ at time t

$$R_{\mathbf{f}}(\xi, t) = \{s_{\mathbf{f}}(t, \xi, u): u \in \mathcal{U}\}.$$

It is well known [27] that if the f_i's are analytic and complete, $R_{\mathbf{f}}(\xi, t)$ is contained in a maximal integral manifold of the distribution

$$\mathcal{S}_{\mathbf{f}} = \text{Lie}\{\text{ad}_{f_0}^k f_i: k \geqq 0, i = 1, \cdots, m\}$$

and its interior with respect to this manifold is nonempty.

For the general case, let $F = \{f_0 + \sum_{i=1}^m \omega_i f_i: \omega \in \Omega\}$. The local flows of the vector fields in F generate a pseudogroup

$$G_F = \{\exp t_1 g_1 \cdots \exp t_k g_k: t_i \in \mathbb{R}, g_i \in F\}.$$

It is known that $N_{\mathbf{f}}(\xi) = \{\phi(\xi): \phi \in G_F\}$ is a C^∞ immersed, connected submanifold of M [26]. For each $\eta \in N_{\mathbf{f}}(\xi)$, let us consider the subset of $N_{\mathbf{f}}(\xi)$

$$N_{\mathbf{f}}^0(\eta) = \left\{\phi(\eta): \phi \in G_F, \sum_{i=1}^k t_i = 0\right\}.$$

In [12] it is claimed that $N_{\mathbf{f}}^0(\eta)$ is the maximal integral manifold of the distribution

$$\Delta_{\mathbf{f}}^0 = \{g_*\phi \circ g^{-1} - h_*\chi \circ h^{-1}, g, h \in G_F, \phi, \chi \in F\}$$

through η. This is not completely true because, if the vector fields of F are not complete, $N_{\mathbf{f}}^0(\eta)$ may be disconnected even in the analytic case (an example can be found in [2]).

In any case the proofs in [12] show that $N_{\mathbf{f}}^0(\eta)$ can be endowed by a structure of immersed, possibly disconnected, submanifold of $N_{\mathbf{f}}(\xi)$, whose tangent space is given at any point by the distribution $\Delta_{\mathbf{f}}^0$. This implies that the connected components of $N_{\mathbf{f}}^0(\eta)$ are maximal integral manifolds of $\Delta_{\mathbf{f}}^0$. Moreover, the codimension of $N_{\mathbf{f}}^0(\eta)$ in $N_{\mathbf{f}}(\xi)$ is either zero or one.

The Lie subalgebra $\mathcal{S}_{\mathbf{f}}$ is always contained in $\Delta_{\mathbf{f}}^0$ and the two distributions coincide if the dimension of $\mathcal{S}_{\mathbf{f}}$ is constant on $N_{\mathbf{f}}(\xi)$. We recall that in the analytic case this last condition is always fulfilled [27]. Let $T \geqq 0$ be such that $R_{\mathbf{f}}(\xi, T)$ is not empty and choose a point $\eta \in R_{\mathbf{f}}(\xi, T)$. $N_{\mathbf{f}}^0(\eta)$ does not depend on the choice of η in $R_{\mathbf{f}}(\xi, T)$. We set $N_{\mathbf{f}}(\xi, T) = N_{\mathbf{f}}^0(\eta)$. By construction $R_{\mathbf{f}}(\xi, T) \subset N_{\mathbf{f}}(\xi, T)$.

DEFINITION 1.1. The relative interior of $R_{\mathbf{f}}(\xi, t)$, $\operatorname{int}_{\mathrm{rel}} R_{\mathbf{f}}(\xi, t)$, is the set of interior points with respect to $N_{\mathbf{f}}(\xi, t)$.

Let $\hat{u} \equiv 0$ be the reference control and let us suppose that the flow $t \to \exp tf_0 \cdot \xi_0 \equiv \hat{x}(t)$ is defined on the compact interval $J = [0, T]$.

We want to give conditions for $(\Sigma_{\mathbf{f}}, \Omega)$ being locally controllable or weakly locally controllable along the reference trajectory $\hat{x}$.

Let us start by defining these properties.

DEFINITION 1.2. $(\Sigma_{\mathbf{f}}, \Omega)$ is *locally controllable* along the trajectory $\hat{x}$ if and only if for each $t \in (0, T]$

$$\hat{x}(t) \in \operatorname{int} R_{\mathbf{f}}(\xi_0, t).$$

Remark 1.1. If $\hat{x}$ is stationary, local controllability along $\hat{x}$ reduces to small time local controllability at the point ξ_0.

DEFINITION 1.3. $(\Sigma_{\mathbf{f}}, \Omega)$ is *weakly locally controllable* along the trajectory $\hat{x}$ if and only if for each $t \in (0, T]$

$$\hat{x}(t) \in \operatorname{int}_{\mathrm{rel}} R_{\mathbf{f}}(\xi_0, t). \tag{1.1}$$

Remark 1.2. $(\Sigma_{\mathbf{f}}, \Omega)$ is locally controllable along $\hat{x}$ if and only if $(\Sigma_{\mathbf{f}}, \Omega)$ is weakly locally controllable along $\hat{x}$ and $\dim N^0_{\mathbf{f}}(\xi_0) = \dim M$.

Remark 1.3. The standard properties of the flow imply that $\hat{x}$ is weakly locally controllable if and only if for each $\tau > 0$ there is $t < \tau$ such that (1.1) holds.

It is known [15] that, in the analytic case, for any $\tau > 0$ there is $t < \tau$ and a trajectory $\gamma : [0, t] \to M$ relative to a piecewise constant control such that $\gamma(t) \in \operatorname{int}_{\mathrm{rel}} R_{\mathbf{f}}(\xi_0, t)$. It is clear that it is possible to prove the local controllability property if a method for bringing $\gamma(t)$ back to the reference trajectory is provided, that is, if a positive integer L and another trajectory $\mu : [0, (L-1)t] \to M$ can be found so that $\mu(0) = \gamma(t)$ and $\mu((L-1)t) = \hat{x}(Lt)$.

The following lemma due to Sussmann [24], explains which kind of conditions we may expect to be sufficient for "bringing $\gamma(t)$ back to the reference trajectory."

LEMMA 1.1. *Let* $\mathbf{f} = \{f_0, \cdots, f_m\}$ *be a family of analytic complete vector fields such that* Lie $\mathbf{f}$ *is a nilpotent Lie algebra. There is an integer* K *such that for each piecewise constant control* u *defined on* $[0, T]$ *with values in a hypercube* $H_\rho = \{(\omega_1, \cdots, \omega_m) \in \mathbb{R}^m : |\omega_i| \leqq \rho < +\infty, i = 1, \cdots, m\}$, *there are*:

(a) $L \leqq K$ *and a piecewise constant control* $\tilde{u}$ *defined on* $[0, (L-1)T]$ *with values in the same hypercube* H_ρ.

(b) *An element* χ *of* $\mathscr{S}_{\mathbf{f}}$ *that is a linear combination of brackets* $\Lambda \in \mathscr{S}_{\mathbf{f}}$ *containing* f_0 *an odd number of times and each* $f_1, \cdots, f_m$ *an even number of times such that*

$$s_{\mathbf{f}}((L-1)T, s_{\mathbf{f}}(T, u, \xi), \tilde{u}) = \exp(LTf_0 + \chi) \cdot \xi.$$

Moreover, χ *may be chosen symmetric with respect to the elements of any subset* $\{f_{i_1}, \cdots, f_{i_r}\}$ *of* $\{f_1, \cdots, f_m\}$, *i.e.,* χ *may be chosen as a fixed element of the automorphisms of* $\mathscr{S}_{\mathbf{f}}$ *generated by* $\mu_{ij} : f_i \mapsto f_j$, $i, j \in \{i_1, \cdots, i_r\} \subseteq \{1, \cdots, m\}$.

Proof. The proof is based on the Campbell-Hausdorff formula and on the symmetries of the control set H_ρ. It is essentially given by the proof of Proposition 5.1 in [24], where the finite group of pseudoautomorphisms of Lie $\mathbf{f}$ is generated by

$$\lambda_i : f_i \mapsto -f_i, \quad i = 1, \cdots, m, \qquad \mu_{ij} : f_i \mapsto f_j, \quad i, j \in \{i_1, \cdots, i_r\}$$

and the so-called "time reversal" automorphism. See [24, § 7]. □

For the stationary case, the above lemma points out as a set of candidates to be "obstruction to small time local controllability" the subspace $\mathscr{B}_{\mathbf{f}}$ spanned by the

brackets that contain f_0 an odd number of times and each $f_1, \cdots, f_m$ an even number of times. In other words, if Lie $\mathbf{f}$ is nilpotent and for each $\chi \in \mathscr{B}_{\mathbf{f}}$, $\chi(\xi_0) = 0$, then $(\Sigma_{\mathbf{f}}, H_\rho)$ is small time locally controllable at ξ_0.

Moreover, it is known that

$$\exp(-LTf_0) \cdot \exp(LTf_0 + \chi) \cdot \xi = \exp \bar{\chi} \cdot \xi \quad \text{with } \bar{\chi} \in \mathscr{S}_{\mathbf{f}} \cap \operatorname{Lie}(f_0, \chi).$$

Therefore in the nonstationary case we are lead to look for the obstructions in $\mathscr{B}_{\mathbf{f}}^* = \mathscr{S}_{\mathbf{f}} \cap \operatorname{Lie}(f_0, \mathscr{B}_{\mathbf{f}})$.

Lemma 1.1 suggests also that the set of "obstructions" could be restricted by selecting a subset of $\{1, \cdots, m\}$. We will do this selection by defining a "weight" on Lie $\mathbf{f}$ and we will give conditions on this new set of obstructions so that a C^∞ system can be considered as an "admissible perturbation" of a nilpotent analytic system for which the obstructions are zero. The meaning of admissible perturbation will be clarified in § 2.

To make the above ideas more precise we need some notation. Let Lie $\mathbf{X}$ be the free Lie algebra on $\mathbb{R}$ generated by the noncommutative indeterminates $\mathbf{X} = \{X_0, \cdots, X_m\}$. We will denote the ideal of Lie $\mathbf{X}$ generated by $X_1, \cdots, X_m$ by $\mathscr{S}$, that is,

$$\mathscr{S} = \operatorname{Lie}\{\operatorname{ad}_{X_0}^k X_i : k \geqq 0, i = 1, \cdots, m\}.$$

Substituting X_i by f_i in any element $\chi \in \operatorname{Lie} \mathbf{X}$, we obtain a vector field that will be denoted by $\chi_{\mathbf{f}}$. For any subset A of Lie $\mathbf{X}$, $A_{\mathbf{f}}$ will denote the subset of Lie $\mathbf{f}$ given by $A_{\mathbf{f}} = \{\chi_{\mathbf{f}} : \chi \in A\}$. By means of a set $\mathbf{l} = (l_0, \cdots, l_m)$ of integers we define a weight on Lie $\mathbf{X}$ that will induce a weight on Lie $\mathbf{f}$.

Let Λ be a bracket in $\mathscr{S}$. We denote the "length" of Λ with respect to X_i (i.e., the number of times that X_i appears in Λ) by $|\Lambda|_i$. If, for example, $\Lambda = [[X_0, X_1], [X_2, [X_1, X_0]]]$, then $|\Lambda|_0 = 2$, $|\Lambda|_1 = 2$, $|\Lambda|_2 = 1$, $|\Lambda|_i = 0$ for all $i > 2$.

The weight of Λ is defined by

$$\|\Lambda\|_{\mathbf{l}} = \sum_{i=0}^{m} l_i |\Lambda|_i, \qquad \|0\|_{\mathbf{l}} = 0$$

and the subspace of $\mathscr{S}_{\mathbf{f}}$ of the elements of weight not greater than i is given by

$$V_i^{\mathbf{l}} = \operatorname{span}\{\Lambda_{\mathbf{f}} : \Lambda \in \mathscr{S}, \|\Lambda\|_{\mathbf{l}} \leqq i\}.$$

An element $\chi \in \mathscr{S}$ is called $\mathbf{l}$-homogeneous if it is a linear combination of brackets with the same weight, which will be called the weight of χ.

Following Sussmann we say that an $\mathbf{l}$-homogeneous element $\chi \in \mathscr{S}$ is $\mathbf{l}$-*neutralized* for $(\Sigma_{\mathbf{f}})$ at ξ_0 if $\chi_{\mathbf{f}}$ is a linear combination at ξ_0 of brackets with less weight. In other words, χ is $\mathbf{l}$-neutralized if there is an $i < \|\chi\|_{\mathbf{l}}$ such that $\chi_{\mathbf{f}}(\xi_0) \in V_i^{\mathbf{l}}(\xi_0)$.

Example 1.1. If $l_0 = l_1 = \cdots = l_m = 1$, $\|\Lambda\|_{\mathbf{l}}$ is the length of the bracket Λ, that is, the number of indeterminates contained in Λ. $V_i^{\mathbf{l}}$ is the linear span of the brackets containing at most i vector fields. $\chi \in \mathscr{S}$ is $\mathbf{l}$-neutralized for $(\Sigma_{\mathbf{f}})$ at ξ_0 if $\chi_{\mathbf{f}}$ is a linear combination at ξ_0 of brackets with less length.

Example 1.2. If $l_0 = 0$, $l_1 = \cdots = l_m = 1$, $\|\Lambda\|_{\mathbf{l}}$ is the number of $X_1, \cdots, X_m$ contained in Λ. $V_i^{\mathbf{l}}$ is the linear span of brackets that contain at most i controlled vector fields. These subspaces are the subspaces S^i introduced by Hermes [11]. $\chi \in \mathscr{S}$ is $\mathbf{l}$-neutralized for $(\Sigma_{\mathbf{f}})$ at ξ_0 if $\chi_{\mathbf{f}}$ is a linear combination at ξ_0 of brackets containing less controlled fields than $\chi_{\mathbf{f}}$.

Example 1.3. If $l_0 = 1$, $l_1 = \cdots = l_m = 0$, $\|\Lambda\|_{\mathbf{l}}$ is the number of X_0 contained in Λ. $V_i^{\mathbf{l}}$ is the linear span of brackets that contain f_0 at most i times. $\chi \in \mathscr{S}$ is $\mathbf{l}$-*neutralized*

for $(\Sigma_{\mathbf{f}})$ at ξ_0 if $\chi_{\mathbf{f}}$ is a linear combination at ξ_0 of brackets containing a lower number of f_0.

To define our candidates to be obstructions we set

$$\mathcal{B} = \text{span}\,\{\Lambda \in \mathcal{S}: |\Lambda|_0 \text{ is odd, } |\Lambda|_i \text{ is even, } i = 1, \cdots, m\},$$

$$\mathcal{B}^{\mathbf{l}}_S = \{\chi \in \mathcal{B}: \chi \text{ is symmetric w.r.t. those } X_i\text{'s, } i \neq 0, \text{ that have the same weight}\}.$$

In other words the elements of $\mathcal{B}^{\mathbf{l}}_S$ are the fixed elements of the automorphisms of $\mathcal{B}$ generated by μ_{ij}: $X_i \mapsto X_j$, for all i, j such that $l_i = l_j i, j \neq 0$.

For example, if $m = 3$, $l_0 = 0$, $l_1 = l_2 = 1$, $l_3 = 3$, then $\text{ad}^2_{X_1} X_0 + \text{ad}^2_{X_2} X_0$ and $\text{ad}^2_{X_3} X_0$ belong to $\mathcal{B}^{\mathbf{l}}_S$, but $\text{ad}^2_{X_1} X_0$ does not.

Note that if $l_1 = \cdots = l_m$ then

$$\mathcal{B}^{\mathbf{l}}_S = \mathcal{B}_S = \{\chi \in \mathcal{B}: \chi \text{ is symmetric w.r.t. } X_1, \cdots, X_m\}$$

and if $l_i \neq l_j$, $i, j = 1, \cdots, m$ then $\mathcal{B}^{\mathbf{l}}_S = \mathcal{B}$.

Finally, we take as obstructions the set

$$\mathcal{B}^*_{\mathbf{l}} = \text{Lie}\,(X_0, \mathcal{B}^{\mathbf{l}}_S) \cap \mathcal{S}.$$

In particular, if $l_1 = \cdots = l_m$ then

$$\mathcal{B}^*_{\mathbf{l}} = \mathcal{B}^*_S = \text{Lie}\,(X_0, \mathcal{B}_S) \cap \mathcal{S}$$

and if $l_i \neq l_j$, $i, j = 1, \cdots, m$ then

$$\mathcal{B}^*_{\mathbf{l}} = \mathcal{B}^* = \text{Lie}\,(X_0, \mathcal{B}) \cap \mathcal{S}.$$

Remark 1.4. In [13] it is proved that there are elements in $\mathcal{B}$ that are not obstructions to small time local controllability. It is an open question which are the real obstructions to small time local controllability.

DEFINITION 1.4. A set $\mathbf{l} = (l_0, \cdots, l_m)$ of integers will be called a *set of admissible weights* for Ω if and only if for each $\omega = (\omega_1, \cdots, \omega_m) \in \Omega$ and each $\varepsilon \in (0, 1)$

$$(\varepsilon^{l_1 - l_0}\omega_1, \cdots, \varepsilon^{l_m - l_0}\omega_m) \in \Omega.$$

Remark 1.5. If Ω is the hypercube $H_\rho = \{(\omega_1, \cdots, \omega_m) \in \mathbb{R}^m: |\omega_i| \leqq \rho < +\infty,\ i = 1, \cdots, m\}$, then $\mathbf{l}$ is a set of admissible weights for H_ρ if and only if $l_0 = \min\{l_0, \cdots, l_m\}$. For such a control set, both weights of Examples 1.1 and 1.2 are admissible but that of Example 1.3 is not.

The weight defined in Example 1.1 is admissible for any Ω; the one defined in Example 1.3 is admissible if and only if Ω is a cone. Note that any set of integers is admissible for $\Omega = \mathbb{R}^m$. Negative weights will be considered in Theorem 2.2.

Below we state the main result on local controllability along a reference trajectory. The assumptions are of two kinds. The first one is a rank condition, and the second one allows us to prove that the obstructions of the approximating nilpotent system defined in § 3 vanish at ξ_0.

THEOREM 1.1. *Let* Ω *be the hypercube* $H_\rho = \{(\omega_1, \cdots, \omega_m) \in \mathbb{R}^m: |\omega_i| \leqq \rho,\ i = 1, \cdots, m\}$, *possibly* $\rho = +\infty$, *i.e.,* $H_\infty = \mathbb{R}^m$. *If*

(a) *the dimension of the manifold* $N^0_{\mathbf{f}}(\xi_0)$ *is equal to the dimension of* $\mathcal{S}_{\mathbf{f}}(\xi_0)$,

(b) *there exists a set of admissible nonnegative weights* $\mathbf{l}$ *such that each* $\mathbf{l}$*-homogeneous element* χ *belonging to the obstruction space* $\mathcal{B}^*_{\mathbf{l}}$ *is* $\mathbf{l}$*-neutralized at* ξ_0, *i.e.,*

$$\forall \chi \in \mathcal{B}^*_{\mathbf{l}} \text{ there is } i < \|\chi\|_{\mathbf{l}} \text{ such that } \chi_{\mathbf{f}}(\xi_0) \in V^{\mathbf{l}}_i(\xi_0)$$

then $(\Sigma_{\mathbf{f}}, \Omega)$ *is weakly locally controllable along* $\hat{x}$.

COROLLARY 1.1. *If* $\dim \mathcal{S}_{\mathbf{f}}(\xi_0) = \dim M$ *and* (b) *holds, then* $(\Sigma_{\mathbf{f}}, \Omega)$ *is locally controllable along* $\hat{x}$.

Taking into account Examples 1.1 and 1.3, the following results can be easily derived.

COROLLARY 1.2 (Hermes [11], Sussmann [24]). *Let* $\dim \mathscr{S}_{\mathbf{f}}(\xi_0) = \dim M$ *and let* Ω *be a neighbourhood of* $0 \in \mathbb{R}^m$. *If each* $\chi_{\mathbf{f}} \in \mathscr{S}_{\mathbf{f}}$ *that contains each controlled vector field an even number of times is a linear combination at* ξ_0 *of brackets with a less number of controlled vector fields, then* $(\Sigma_{\mathbf{f}}, \Omega)$ *is locally controllable along* $\hat{x}$.

Proof. Since each element of $\mathscr{B}^*$ contains each controlled vector field an even number of times, each obstruction is neutralized by means of the weights $l_0 = 0$, $l_1 = \cdots = l_m = 1$.

COROLLARY 1.3. *If the dimension of* Lie $(f_1, \cdots, f_m)$ *at* ξ_0 *is equal to the dimension of* M, *then* $(\Sigma_{\mathbf{f}}, \mathbb{R}^m)$ *is locally controllable along* $\hat{x}$.

Proof. If $l_0 = 1$ and $l_1 = \cdots = l_m = 0$ then each obstruction has weight at least one so that it is $\mathbf{l}$-neutralized because it is a linear combination at ξ_0 of brackets of weight zero.

Example 1.4. Let $M = \mathbb{R}^3$, $\xi_0 = (0, 0, 0)$, $\Omega = \{\omega \in \mathbb{R}: |\omega| \leqq 1\}$, and

$$f_0 = \frac{\partial}{\partial x}, \qquad f_1 = \frac{\partial}{\partial x} + x\frac{\partial}{\partial y} + (x^3 y + y^2)\frac{\partial}{\partial z}.$$

The significant Lie brackets are:

$$[f_0, f_1] = \frac{\partial}{\partial y} + 3x^2 y \frac{\partial}{\partial z}, \qquad \mathrm{ad}^2_{f_0} f_1 = 6xy\frac{\partial}{\partial z},$$

$$\mathrm{ad}^2_{f_1} f_0 = -(2x^3 + 6xy - 2y)\frac{\partial}{\partial z}, \qquad [f_0, \mathrm{ad}^2_{f_1} f_0] = -(6x^2 + 6y)\frac{\partial}{\partial z},$$

$$\mathrm{ad}^2_{f_0}\mathrm{ad}^2_{f_1} f_0 = -12x\frac{\partial}{\partial z}, \quad [\mathrm{ad}_{f_0} f_1, \mathrm{ad}^2_{f_0} f_1] = 6x\frac{\partial}{\partial z}, \quad [\mathrm{ad}_{f_0} f_1, \mathrm{ad}^2_{f_1} f_0] = -(2 - 6x)\frac{\partial}{\partial z}.$$

Therefore, if we choose $l_0 = 2$, $l_1 = 3$, $V^1_{13} = \mathbb{R}^3$ and the brackets in $\mathscr{B}^*$ that have a weight less than or equal to 13 are $\mathrm{ad}^i_{f_0}\mathrm{ad}^2_{f_1 f_0}$, $i = 0, 1, 2$, and $[\mathrm{ad}_{f_0} f_1, \mathrm{ad}^2_{f_0} f_1]$. They are zero at ξ_0, so that they are neutralized. Therefore Theorem 1.1 implies that $(t, 0, 0) \in \operatorname{int} R_{\mathbf{f}}(\xi_0, t)$ for each t. Note that for this system Theorem 1.1 does not apply with $\mathbf{l} = (0, 1)$. In fact, $\mathrm{ad}^3_{f_0}\mathrm{ad}^2_{f_1} f_0 = -12(\partial/\partial z)$, so that $S^2(\xi_0)$ is not contained in $S^1(\xi_0)$ and the obstruction $\mathrm{ad}^3_{f_0}\mathrm{ad}^2_{f_1} f_0$ is not $\mathbf{l}$-neutralized.

Remark 1.6. Some authors, for example, Petrov [18] and Goncalves [8], give a more restrictive definition of local controllability at ξ_0. Namely, they also require that the restriction of the system to each neighbourhood of ξ_0 is locally controllable. An analogous definition of local controllability along a reference trajectory can be given. Conditions (a) and (b) of Theorem 1.1 are conditions on the germs of the vector fields $f_0, \cdots, f_m$ at ξ_0, so that they guarantee that each restriction of $(\Sigma_{\mathbf{f}}, H_\rho)$ to any neighbourhood of ξ_0 is locally controllable along the reference trajectory. Note that the hypotheses of the theorem guarantee that for each neighbourhood V of $\hat{x}([0, T])$, the restriction of the system to V is locally controllable along $\hat{x}$.

In [24] Sussmann states very general sufficient conditions of small time local controllability that generalize many of those known before, but an example in [1] shows that they are no longer sufficient in the nonstationary case. To compare Theorem 1.1 with the results in [24], let us recall the main result stated there.

Let $f_0(\xi_0) = 0$, if there is a set of admissible nonnegative weights $\mathbf{l}$ such that

(i) Lie $(f_0, \cdots, f_m)$ has full rank at ξ_0,

(ii) each element in $\mathscr{B}^{\mathbf{l}}_S$ is $\mathbf{l}$-neutralized at ξ_0,

then (Σ, H_ρ) is small time locally controllable at ξ_0.

The example in [1] shows that the $\mathbf{l}$-neutralization of the elements of $\mathscr{B}^{\mathbf{l}}_S$ is not sufficient to get local controllability in the nonstationary case. But the following lemma,

whose proof can be found in [23], shows that if each element of $\mathscr{B}_S^{\mathbf{l}}$ is $\mathbf{l}$-neutralized at each point of the reference trajectory, then the same holds for each element of $\mathscr{B}_{\mathbf{l}}^*$.

LEMMA 1.2. *If* $\dim \mathscr{V}_i^{\mathbf{l}}$ *is constant along the trajectory* $\hat{x}$ *and each element of* $\mathscr{B}_S^{\mathbf{l}}$ *is* $\mathbf{l}$*-neutralized at each point of the reference trajectory, then the same holds for each element in* $\mathscr{B}_{\mathbf{l}}^*$.

As a consequence we get that if $\hat{x}$ is stationary, then the elements of $\mathscr{B}_{\mathbf{l}}^*$ are $\mathbf{l}$-neutralized at ξ_0 if and only if the ones of $\mathscr{B}_S^{\mathbf{l}}$ are $\mathbf{l}$-neutralized. Therefore Theorem 1.1 includes the result of Sussmann quoted before, and proves that small time local controllability can be viewed as a particular case of local controllability along a reference trajectory.

Let us end this section with an application of the Theorem 1.1 to the case of unbounded controls.

COROLLARY 1.4. *Let* $W_i = \operatorname{span}\{\chi_{\mathbf{f}}: \chi \in \mathscr{S},\ |\chi|_0 = 1,\ \sum_{j=1}^m |\chi|_j \leqq i\}$. *If*

(a) $\mathscr{S}_{\mathbf{f}}(\xi_0)$ *is spanned by the brackets which contain* f_0 *at most once,*

(b) $\mathscr{B}_S(\xi_0) \cap W_{2k}(\xi_0) \subset W_{2k-1}(\xi_0) + \operatorname{Lie}(f_1, \cdots, f_m)(\xi_0)$,

then $(\Sigma_{\mathbf{f}}, \mathbb{R}^m)$ *is weakly locally controllable along the trajectory* $\hat{x}$.

Proof. Set $Y_0 = \operatorname{Lie}\{f_1, \cdots, f_m\}$ and $Y_1 = \operatorname{span}\{\chi_{\mathbf{f}}: \chi \in \operatorname{Lie} X, |\chi|_0 = 1\}$. Let $\chi_{\mathbf{f}}^1(\xi_0), \cdots, \chi_{\mathbf{f}}^r(\xi_0)$ be a basis for $Y_1(\xi_0) + Y_0(\xi_0)$ and let

$$h = \max\left\{\sum_{j=0}^m |\chi^i|_j : i = 1, \cdots, r\right\}.$$

If $\mathbf{l} = (h, 1, \cdots, 1)$ then $Y_1(\xi_0) + Y_0(\xi_0) = \mathscr{S}_{\mathbf{f}}(\xi_0) = V_{2h-1}^{\mathbf{l}}(\xi_0)$, $Y_0(\xi_0) \subseteq V_h^{\mathbf{l}}(\xi_0)$, and each $\Lambda \in \mathscr{B}_{\mathbf{l}}^* = \mathscr{B}_S^*$ is such that $\|\Lambda\|_{\mathbf{l}} \geqq h+2$. Therefore if $\Lambda \in \mathscr{B}_{\mathbf{l}}^*$ and $|\Lambda|_0 = 1$, then $\Lambda \in \mathscr{B}_S$ and condition (b) implies that Λ is $\mathbf{l}$-neutralized. Otherwise $|\Lambda|_0 \geqq 2$ so that $\|\Lambda\|_{\mathbf{l}} > 2h > 2h-1$. □

In [8] a sufficient condition of local controllability for a system with unbounded controls is given. They can be compared in the following way. For the scalar input case Corollary 1.4 is a stronger result. Namely, condition (b) is replaced by $W_{2k} = W_{2k-1}$. For the multi-input case the results are different and none of them is a consequence of the other as the following examples show.

Example 1.5. Let $M = \mathbb{R}^3$, $\xi_0 = (0, 0, 0)$, $\Omega = \mathbb{R}^2$, and

$$(\Sigma_{\mathbf{f}}) \qquad \dot{x} = u_1, \quad \dot{y} = u_2 x^6 + x^2, \quad \dot{z} = x^2 y$$

$Y_0(\xi_0) + W_3(\xi_0) = \mathbb{R}^3$ and $W_2(\xi_0) \subset Y_0(\xi_0)$, so that $(\Sigma_{\mathbf{f}})$ satisfies the hypotheses of Corollary 1.4, but it does not satisfy the hypotheses in [8].

Note that the above system is not small time locally controllable if Ω is bounded because, in this case, $\dot{y}(t)$ is positive for small t.

Example 1.6. Let $M = \mathbb{R}^4$, $\xi_0(0, 0, 0, 0)$, $\Omega = \mathbb{R}^2$, and

$$(\Sigma_{\mathbf{f}}) \qquad \begin{aligned} \dot{x} &= u_1, & \dot{y} &= u_2, \\ \dot{z} &= x^2 y^2 + x^5, & \dot{w} &= x^2 y^2 + y^5. \end{aligned}$$

$(\Sigma_{\mathbf{f}})$ satisfies the hypotheses in [8], but satisfies neither the hypotheses of Corollary 1.4 nor those of Theorem 1.1. In fact, the obstruction $\operatorname{ad}_{f_2}^2 \operatorname{ad}_{f_1}^2 f_0 + \operatorname{ad}_{f_1}^2 \operatorname{ad}_{f_2}^2 f_0$ is equal to $2/5! \operatorname{ad}_{f_1}^5 f_0 + 2/5! \operatorname{ad}_{f_2}^2 f_0$ so that it cannot be neutralized by any set of nonnegative weights.

2. Graded approximating systems. Let $\mathbf{x} = (x^1, \cdots, x^n)$ be a chart on a neighbourhood U of $\xi_0 \in M$ such that $\mathbf{x}(\xi_0) = 0$ and $\mathbf{x}(U)$ is a ball centred at zero. Let $\mathbf{w} = (w^1, \cdots, w^n)$ be a set of positive integers; by means of the couple $(\mathbf{x}, \mathbf{w})$ we will give a graded structure to U that will be called a local graded structure at ξ_0 and will be denoted by $(\mathbf{x}, U, \boldsymbol{w})$.

A graded structure will be named trivial if $w^1 = \cdots = w^n = 1$.

The subalgebra $\mathcal{P}$ of $\mathcal{F}(U)$ generated by the coordinate functions $x^1, \cdots, x^n$ will be called the algebra of polynomials (induced by the chart).

A vector field f will be called polynomial if $f \cdot \varphi \in \mathcal{P}$ for each $\varphi \in \mathcal{P}$, i.e., if $f^i \in \mathcal{P}$, $i = 1, \cdots, n$.

In [9] a graded structure on $\mathbb{R}^n$ is defined and its properties are stated. We refer to it for the details. Here we define a graded structure on U slightly modifying the definitions of [9]. Namely, we give a graded structure to $\mathbb{R}^n$ and we transfer it to U by means of the chart $\mathbf{x}$.

The graded structure is given by the "dilations," i.e., maps δ_ε on U with values in U defined by

$$x^i \circ \delta_\varepsilon = \varepsilon^{w^i} x^i, \qquad i = 1, \cdots, n, \quad \varepsilon \in \mathbb{R}.$$

If $|\varepsilon| > 1$ the dilation δ_ε will be defined in a suitable neighbourhood of ξ_0 contained in U, but nevertheless we will write $\delta_\varepsilon : U \to U$ understanding "locally defined."

A polynomial $\varphi \in \mathcal{P}$ is called *homogeneous of weight a* if $\varphi \circ \delta_\varepsilon = \varepsilon^a \varphi$. The set of polynomials of weight a will be denoted by $\mathcal{P}(a)$ and we get a gradation $\mathcal{P} = \bigoplus_{a \geqq 0} \mathcal{P}(a)$ of the algebra of polynomials.

The *weight* $\mathcal{W}(\varphi)$ *of a polynomial* φ is defined by

$$\mathcal{W}(\varphi) \leqq k \quad \text{iff } \varphi \in \bigoplus_{a=0}^{k} \mathcal{P}(a).$$

In other words, we give the "weight" w^i to x^i. As a consequence, for each multi-index $\alpha = (\alpha_1, \cdots, \alpha_n)$, the weight of the monomial $x^\alpha \equiv (x^1)^{\alpha_1} \cdots (x^n)^{\alpha_n}$ is defined by $\mathcal{W}(x^\alpha) = \sum_{i=1}^n w^i \alpha_i$. Hence the weight of a polynomial φ is the greatest weight of the monomials contained in it. If the graded structure is the trivial one, then $\mathcal{W}(\varphi)$ is the degree of the polynomial.

A polynomial vector field f is called *homogeneous of weight i* if for each $\varphi \in \mathcal{P}(a)$, $f \cdot \varphi \in \mathcal{P}(a - j)$ (if $a < 0$, $\mathcal{P}(a)$ is understood to be equal to $\{0\}$). Roughly speaking if the weight of f is j then f "substracts" weight j from the functions. Therefore, in terms of components, f is homogeneous of weight j if and only if $f^i \in \mathcal{P}(w^i - j)$. For example, if $\mathbf{w} = (1, 2, 3)$, the vector field $x^1(\partial/\partial x^2) + [(x^1)^2 - 4x^2]\,\partial/\partial x^3$ is homogeneous of weight one. Let us remark that in a trivial graded structure the homogeneous vector fields with positive weight are the constant vector fields. In fact, the weight of a homogeneous vector field in a trivial graded structure is equal to one minus the degree of the field.

The set of homogeneous polynomials vector fields of weight j will be denoted by $\mathcal{V}(j)$. Let us remark that $\partial/\partial x^i \in \mathcal{V}(w^i)$ and that $\mathcal{V}(j) = \{0\}$ if $j > \max\{w^1, \cdots, w^n\}$. Moreover, in terms of dilations, $f \in \mathcal{V}(j)$ if and only if $(\delta_\varepsilon)_* f \circ \delta_\varepsilon^{-1} = \varepsilon^j f$.

The *weight* $\mathcal{W}(f)$ *of a polynomial vector field* is defined by

$$\mathcal{W}(f) \leqq j \quad \text{iff } f \in \bigoplus_{k \geqq j} \mathcal{V}(k).$$

It is easy to see that if $f \in \mathcal{V}(i)$ and $g \in \mathcal{V}(j)$, then $[f, g] \in \mathcal{V}(i + j)$, therefore the set $\mathcal{N}^0 = \bigoplus_{i \geqq 0} \mathcal{V}(i)$ of the polynomial vector fields of weight zero is a subalgebra of the Lie algebra of the polynomial vector fields and the set $\mathcal{N} = \bigoplus_{i \geqq 1} \mathcal{V}(i)$ is a nilpotent ideal of $\mathcal{N}^0$.

The *graded order* $\mathcal{O}(\varphi)$ of a polynomial φ is defined by

$$\mathcal{O}(\varphi) \geqq i \quad \text{iff } \varphi \in \bigoplus_{j \geqq i} \mathcal{P}(j).$$

The definition of graded order can be extended in an obvious way to the elements of $\mathcal{F}(U)$. Namely, $\mathcal{O}(\varphi) \geqq i$, if each Taylor approximation at ξ_0 (in the chart $\mathbf{x}$) of φ has graded order $\geqq i$.

The *graded order* $\mathcal{O}(S)$ of a differential operator S is defined by saying that

$$\mathcal{O}(S) \leqq j \quad \text{iff } \mathcal{O}(S \cdot \varphi) \geqq \mathcal{O}(\varphi) - j.$$

In particular, if $f \in \mathcal{V}(U)$, then

$$\mathcal{O}(f) \leqq j \quad \text{iff } \mathcal{O}(f^i) \geqq w^i - j, \quad i = 1, \cdots, n.$$

Example 2.1. Let $\mathbf{w} = (1, 2)$, if $f(x^1, x^2) = \sin(x^1 + x^2)\, \partial/\partial x^1 + \cos(x^1 + x^2)\, \partial/\partial x^2$; then

$$f(x^1, x^2) = (x^1 + x^2 + \cdots) \frac{\partial}{\partial x_1} + \left(1 + \frac{(x_1 + x_2)^2}{2} + \cdots\right) \frac{\partial}{\partial x^2},$$

so that $\mathcal{O}(f) = 2$.

It is easy to see that the following properties hold:

(P1) $\forall S_1, S_2 \in \mathcal{D}(U)$, $\mathcal{O}(S_i) \leqq j_i$, $i = 1, 2 \Rightarrow \mathcal{O}(S_1 \cdot S_2) \leqq j_1 + j_2$ so that $\forall g_1, g_2 \in \mathcal{V}(U)$, $\mathcal{O}(g_i) \leqq j_i$, $i = 1, 2 \Rightarrow \mathcal{O}([g_1, g_2]) \leqq j_1 + j_2$

(P2) For each $\varphi \in \mathcal{F}(U)$ and each integer $k \geqq 0$, there is a unique polynomial $\varphi_{(k)}$ of weight k such that $\mathcal{O}(\varphi - \varphi_{(k)}) \geqq k + 1$. $\varphi_{(k)}$ is called *the graded approximation of weight k* of φ and it is the sum of the polynomials of weight less than or equal to k in the formal Taylor expansion of φ at ξ_0. $\varphi_{(k)}$ coincides also with the graded approximation of weight k of each Taylor approximation at ξ_0 of order greater than k.

(P3) For each $f \in \mathcal{V}(U)$ and each integer $k \leqq \max\{w^1, \cdots, w^n\}$, there is a polynomial vector field $f_{(k)}$ of weight k such that $\mathcal{O}(f - f_{(k)}) \leqq k - 1$. $f_{(k)}$ is called *the graded approximation of weight k* of f and it is the sum of the homogeneous vector fields of weight greater than or equal to k in the formal Taylor expansion of f at ξ_0.

For example if f and $\mathbf{w}$ are as in the Example 2.1, then

$$f_{(0)}(x^1, x^2) = x^1 \frac{\partial}{\partial x^1} + \left(1 - \frac{1}{2}(x^1)^2\right) \frac{\partial}{\partial x^2}.$$

(P4) If $\varphi \in \mathcal{F}(U)$, $\mathcal{O}(\varphi) \geqq s$ if and only if the function defined on $(-1, 1) \times U$ by

$$(\varepsilon, \xi) \to \begin{cases} \varepsilon^{-s} \varphi(\delta_\varepsilon(\xi)), & \varepsilon \neq 0, \\ \varphi_{(s)}(\xi), & \varepsilon = 0 \end{cases}$$

is a C^∞ map.

(P5) If $f \in \mathcal{V}(U)$, $\mathcal{O}(f) \leqq s$ if and only if the map defined on $(-1, 1) \times U$ by

$$(\varepsilon, \xi) \to \begin{cases} \varepsilon^s (\delta_\varepsilon^{-1})_* f(\delta_\varepsilon(\xi)), & \varepsilon \neq 0, \\ f_{(s)}(\xi), & \varepsilon = 0 \end{cases}$$

is a C^∞ map.

In the sequel we will use the notation introduced in § 1.

Let a local graded structure $(\mathbf{x}, U, \mathbf{w})$ of M at ξ_0 be fixed. Let $\mathbf{f} = \{f_0, \cdots, f_m\}$ and $\mathbf{l} = (l_0, \cdots, l_m)$ be a set of integers such that $l_i \geqq \mathcal{O}(f_i)$. The graded approximation of weight l_i of f_i will be denoted by $\hat{f}_i$; it belongs to $\mathcal{V}(l_i)$, $i = 0, \cdots, m$. Note that $\hat{f}_i = 0$ if and only if $l_i > \mathcal{O}(f_i)$. Set $-\mathbf{f} = \{\hat{\mathbf{f}}_0, \cdots, \hat{\mathbf{f}}_m\}$.

DEFINITION 2.1. The system $(\Sigma_{\hat{\mathbf{f}}})$ defined on U is named a *graded approximation* of the system $(\Sigma_{\mathbf{f}})$.

Examples of graded approximations of a control system can be found in Example 3.1.

Remark 2.1. From $\mathscr{W}(\hat{f}_i^j) = w^j - l_i$, it follows that if $l_i = 0$, then $\hat{f}_i^j$ depends linearly on the x^k's with $w^k = w^j$ and does not depend on the x^k's with $w^k > w^j$. If $l_i > 0$, then $\hat{f}_i^j$ does not depend on any x^k with $w^k \geqq w^j$. In other words, if each l_i is nonnegative, then the system $(\Sigma_{\hat{\mathbf{f}}})$ is a cascade of linear systems and integrators; if each l_i is positive it is a cascade of integrators.

Remark 2.2. $\mathscr{S}_{\hat{\mathbf{f}}}$ is a distribution spanned by homogeneous vector fields. Hence, possibly after a linear change of coordinates that does not affect the graded structure, we can suppose that

$$\mathscr{S}_{\hat{\mathbf{f}}}(\xi_0) = \operatorname{span}\left\{\frac{\partial}{\partial x^1}, \cdots, \frac{\partial}{\partial x^r}\right\},$$

r being the dimension of $\mathscr{S}_{\hat{\mathbf{f}}}(\xi_0)$.

LEMMA 2.1. *If Λ is any bracket in $\mathscr{S}$*

(a) $\mathscr{O}(\Lambda_{\mathbf{f}} - \Lambda_{\hat{\mathbf{f}}}) < \|\Lambda\|_{\mathbf{l}}$,

(b) $\mathscr{S}_{\mathbf{f}}(\xi_0) + \operatorname{span}\{\partial/\partial x^{r+1}, \cdots, \partial/\partial x^n\} = T_{\xi_0}M$.

Proof. Part (a) is obvious by definitions.

(b) Let $\Lambda_1, \cdots, \Lambda_r$ be brackets in $\mathscr{S}$ such that

$$\operatorname{span}\{\Lambda_{1\hat{\mathbf{f}}}(\xi_0), \cdots, \Lambda_{r\hat{\mathbf{f}}}(\xi_0)\} = \mathscr{S}_{\hat{\mathbf{f}}}(\xi_0).$$

When we use (a) it is not difficult to see that

$$\Lambda_{1\mathbf{f}}(\xi_0), \cdots, \Lambda_{r\mathbf{f}}(\xi_0), \frac{\partial}{\partial x^{r+1}}(\xi_0), \cdots, \frac{\partial}{\partial x^n}(\xi_0)$$

are linearly independent. □

To the set of integers $\mathbf{l} = (l_0, \cdots, l_m)$ we can associate a "one-parameter family of variations" of the null control in $L^1([0, T], \mathbb{R}^m)$, namely, a map

$$\delta : (-1, 1) \times L^1([0, T], \mathbb{R}^m) \to L^1(\mathbb{R}, \mathbb{R}^m)$$

defined by

$$\delta(\varepsilon, u) \equiv \delta_\varepsilon u \equiv u_\varepsilon : t \mapsto \begin{cases} (\varepsilon^{l_1 - l_0} u_1(t/\varepsilon^{l_0}), \cdots, \varepsilon^{l_m - l_0} u_m(t/\varepsilon^{l_0})) & \text{if } t \in [0, \varepsilon^{l_0} T], \\ 0 \quad \text{otherwise.} \end{cases}$$

The following results are devoted to relating the solutions of $(\Sigma_{\hat{\mathbf{f}}})$ and $(\Sigma_{\mathbf{f}})$ through the dilations defined above. We can say that the solution of $(\Sigma_{\hat{\mathbf{f}}})$ relative to u gives the principal part with respect to ε of the trajectory's variation induced by $\delta_\varepsilon u$ along $\hat{x}$.

THEOREM 2.1. *Let $\bar{t} \in [0, T]$, $\bar{\xi} \in U$ and $\bar{u} \in L^1([0, \bar{t}], \mathbb{R}^m)$ be such that $s_{\hat{\mathbf{f}}}(\cdot, \bar{\xi}, \bar{u})$ is defined in $[0, \bar{t}]$. There exists a neighbourhood $V(\bar{\xi})$ of $\bar{\xi}$ in U, a neighbourhood $V(\bar{u})$ of $\bar{u}$ in $L^1([0, \bar{t}], \mathbb{R}^m)$, and $\bar{\varepsilon} > 0$ such that for all $\xi \in V(\bar{\xi})$, $u \in V(\bar{u})$, and $\varepsilon \in (-\bar{\varepsilon}, \bar{\varepsilon})$, $\varepsilon \neq 0$, $\delta_\varepsilon^{-1} s_{\mathbf{f}}(\cdot, \delta_\varepsilon \xi, u_\varepsilon)$ is defined in $[0, \varepsilon^{l_0}\bar{t}]$ (on $[\varepsilon^{l_0}\bar{t}, 0]$ if $\varepsilon < 0$). Moreover,*

(a) *The map $S : V(\bar{\xi}) \times V(\bar{u}) \times (-\bar{\varepsilon}, \bar{\varepsilon}) \to U$ defined by*

$$S(\xi, u, \varepsilon) = \begin{cases} \delta_\varepsilon^{-1} s_{\mathbf{f}}(\varepsilon^{l_0}\bar{t}, \delta_\varepsilon \xi, u_\varepsilon), & \varepsilon \neq 0, \\ s_{\hat{\mathbf{f}}}(\bar{t}, \xi, u), & \varepsilon = 0 \end{cases}$$

is a C^∞ map;

(b) $s_{\hat{\mathbf{f}}}(\varepsilon^{l_0}\bar{t}, \delta_\varepsilon \xi, u_\varepsilon) = \delta_\varepsilon s_{\hat{\mathbf{f}}}(\bar{t}, \xi, u)$.

Proof. It is not restrictive to assume that U is an open neighbourhood of zero in $\mathbb{R}^n$. Let $\varepsilon \neq 0$, $\xi \in U$, and $u \in L^1([0, \bar{t}], \mathbb{R}^m)$; the map $t \mapsto \delta_\varepsilon^{-1} s_{\mathbf{f}}(\varepsilon^{l_0} t, \delta_\varepsilon \xi, u_\varepsilon)$ is the solution of the differential equation on U

$$\dot{x} = \varepsilon^{l_0}(\delta_\varepsilon^{-1})_* f_0(\delta_\varepsilon x) + \sum_{i=1}^{m} u_i \varepsilon^{l_i}(\delta_\varepsilon^{-1})_* f_i(\delta_\varepsilon x),$$

which at time zero is equal to ξ. Hence it is defined implicitly by the equation

$$\psi(t) = \xi + \int_0^t \left[F_0(\varepsilon, \psi(s)) + \sum_{i=1}^{m} u_i(s) F_i(\varepsilon, \psi(s)) \right] ds,$$

with

$$F_j(\varepsilon, x) = \begin{cases} \varepsilon^{l_j}(\delta_\varepsilon^{-1})_* f_j(\delta_\varepsilon x), & \varepsilon \neq 0, \\ \hat{f}_j(x), & \varepsilon = 0. \end{cases}$$

Let $G: U \times (-1, 1) \times L^1([0, \bar{t}], \mathbb{R}^m) \times C^0([0, \bar{t}], U) \to C^0([0, \bar{t}], \mathbb{R}^n)$ be the map defined by

$$G(\xi, \varepsilon, u, \psi)(t) = \psi(t) - \xi - \int_0^t \left[F_0(\varepsilon, \psi(s)) + \sum_{i=1}^{m} u_i(s) F_i(\varepsilon, \psi(s)) \right] ds.$$

By property (P5) the F_j's are C^∞ on $(-1, 1) \times U$; hence G is a C^∞ map (see [5]). $F_j(0, x) = \hat{f}_j(x)$, so that $G(\bar{\xi}, 0, \bar{u}, s_{\mathbf{f}}(\bar{\xi}, \bar{u}, \cdot)) = 0$. Moreover, in any point the derivative of G with respect to ψ is a linear homeomorphism of $C^0([0, \bar{t}], \mathbb{R}^n)$ onto $C^0([0, \bar{t}], \mathbb{R}^n)$ [16] so that the implicit function theorem implies (a). Statement (b) follows by $\hat{f}_i \in V(l_i)$. □

COROLLARY 2.1. *Let $\xi \in U$ and $u \in L^1([0, \bar{t}], \mathbb{R}^m)$ be such that $s_{\mathbf{f}}(\cdot, \xi, u)$ is defined in $[0, \bar{t}]$; then $s_{\mathbf{f}}(\varepsilon^{l_0}\bar{t}, \delta_\varepsilon \xi, u_\varepsilon)$ is defined for ε sufficiently small and*

$$\mathbf{x}(s_{\mathbf{f}}(\varepsilon^{l_0}\bar{t}, \delta_\varepsilon \xi, u_\varepsilon)) = \mathbf{x}(\delta_\varepsilon(s_{\hat{\mathbf{f}}}(\bar{t}, \xi, u))) + \varepsilon \mathbf{x}(\delta_\varepsilon h(\varepsilon, \xi, u)) \tag{2.1}$$

with $\mathbf{x}(h(\varepsilon, \xi, u))$ bounded on compact sets.

Remark 2.3. In [4] Bressan defines an approximating system by means of a graded structure related to the family $\mathbf{f}$. Moreover, he compares the system with its approximation through a linear rescaling of time and space coordinates. The approximating system defined in [4] is a graded approximating system $(\Sigma_{\hat{\mathbf{f}}})$ defined above. Therefore Theorem 2.1 generalizes the result in [4].

Remark 2.4. The L^1 norm of $\delta_\varepsilon u$ is given by

$$\|\delta_\varepsilon u\|_{L^1} = \sum_{i=1}^{m} \varepsilon^{l_i} \int_0^{\bar{t}} |u_i| \, ds.$$

Hence if one of the l_i is negative, $\|\delta_\varepsilon u\|_{L^1} \to +\infty$ as $\varepsilon \to 0$, and it could be surprising that nevertheless the relative solution goes to ξ_0. The explanation for this is that $l_i < 0$ implies that $f_i \to 0$ very fast near ξ_0.

The following theorem provides the main approximation result. It states that if there is a graded structure at ξ_0 such that the vector fields f_i's are homogeneous, then the local controllability property is stable under perturbations of the vector fields f_i's that do not affect the graded order.

THEOREM 2.2. *Let us suppose that*

(a) $\mathbf{l}$ *is a set of admissible weights such that* $l_0 \geqq 0$, $l_i \geqq \mathcal{O}(f_i)$, $i = 0, 1, \cdots, m$.

(b) $\dim \mathcal{S}_{\hat{\mathbf{f}}}(\xi_0) = \dim N_{\hat{\mathbf{f}}}^0(\xi_0) = r$.

(c) $(\Sigma_{\hat{\mathbf{f}}}, \Omega)$ *is weakly locally controllable along* $t \mapsto \exp t\hat{f}_0 \cdot \xi_0$.

Then, $(\Sigma_{\mathbf{f}}, \Omega)$ *is weakly locally controllable along the trajectory* $t \mapsto \exp tf_0 \cdot \xi_0$.

Proof. Without loss of generality we can suppose that U is an open neighbourhood of ξ_0 in $\mathbb{R}^n$, and $\xi_0 = 0$.

Taking into account Lemma 2.1, the assumption (b) implies that the tangent space at ξ_0 to $N^0_{\mathbf{f}}(\xi_0)$ is $\mathscr{S}_{\mathbf{f}}(\xi_0)$. We will consider $\mathbb{R}^n$ as $\mathbb{R}^r \times \mathbb{R}^{n-r}$ so that, again by Lemma 2.1, the point on $\mathbb{R}^r$ can be seen as coordinates both on $N^0_{\hat{\mathbf{f}}}(\xi_0)$ and $N^0_{\mathbf{f}}(\xi_0)$. To be more precise there are:

(i) Open neighbourhoods V, W, W' of ξ_0 in $\mathbb{R}^r$, $N^0_{\mathbf{f}}(\xi_0)$ and $N^0_{\hat{\mathbf{f}}}(\xi_0)$, respectively;

(ii) C^∞ homeomorphisms $\phi : V \to W$ and $\phi' : V \to W'$ such that ϕ^{-1} and ϕ'^{-1} are equal to the canonical projection π on $\mathbb{R}^r$ restricted to W and W', respectively.

Let $\bar{t} > 0$ be such that $\exp t\hat{f}_0 \cdot \xi_0 \in \operatorname{int}_{\mathrm{rel}} R_{\hat{\mathbf{f}}}(\xi_0, t)$ for any t belonging to the interval $[t', \bar{t}]$; Lemma A in the Appendix implies that there exists a neighbourhood W'' of ξ_0 in $N^0_{\hat{\mathbf{f}}}(\xi_0)$ and a continuous map $\mu : W'' \to \mathscr{U} \cap L^1([0, \bar{t}], \Omega)$ such that

$$\exp(-\bar{t}\hat{f}_0) \cdot s_{\hat{\mathbf{f}}}(\bar{t}, \xi_0, \mu(w)) = w \quad \forall w \in W''.$$

Without loss of generality suppose $W'' = W'$. By possibly restricting W', by Theorem 2.1 and the definition of the topology on $N^0_{\hat{\mathbf{f}}}(\xi_0)$, there is $\bar{\varepsilon}$ such that

$$\exp(-\varepsilon^{l_0}\bar{t}f_0) \cdot s_{\mathbf{f}}(\varepsilon^{l_0}\bar{t}, \xi_0, \delta_\varepsilon \mu(w)) \in W \quad \forall w \in W' \quad \forall \varepsilon \in [0, \bar{\varepsilon}].$$

Let $G : W' \times [0, \bar{\varepsilon}] \to W$ be defined by

$$G(w, \varepsilon) = \exp(-\varepsilon^{l_0}\bar{t}f_0) \cdot s_{\mathbf{f}}(\varepsilon^{l_0}\bar{t}, \xi_0, \delta_\varepsilon \mu(w)).$$

By Theorem 2.1 we get for any $v \in V$

$$\pi\delta_\varepsilon^{-1} G(\phi'(v), \varepsilon) = \pi(\phi'(v) + O(\varepsilon)) = v + O(\varepsilon).$$

Standard arguments in degree theory imply that there is $\varepsilon > 0$ such that $\pi\delta_\varepsilon^{-1} G(\phi'(V), \varepsilon)$ contains an open neighbourhood of ξ_0 in V. But $\pi\delta_\varepsilon^{-1} = \delta_\varepsilon^{-1}\pi$, hence $G(W', \varepsilon)$ contains an open neighbourhood of ξ_0 in $N^0_{\mathbf{f}}(\xi_0)$. $\mathbf{I}$ being an admissible set of weights and $\varepsilon^{l_0}\bar{t}$ less or equal to $\bar{t}$, the theorem is proved. □

COROLLARY 2.2. *If* $(\Sigma_{\hat{\mathbf{f}}})$ *is locally controllable along the trajectory* $t \mapsto \exp t\hat{f}_0 \cdot \xi_0$, *then* $(\Sigma_{\mathbf{f}})$ *is locally controllable along the trajectory* $t \mapsto \exp tf_0 \cdot \xi_0$.

Remark 2.5. In the proof of the above theorem, the assumption $l_0 \geqq 0$ is useful only in proving that the reference trajectory is in the relative interior of the reachable set at a time not greater than $\bar{t}$. Therefore the same proof gives the result stated below.

Let us suppose

(a) $\mathbf{I}$ is a set of admissible weights such that $l_i \geqq \mathscr{O}(f_i)$, $i = 0, 1, \cdots, m$,

(b) $\dim \mathscr{S}_{\hat{\mathbf{f}}}(\xi_0) = \dim N^0_{\mathbf{f}}(\xi_0) = r$,

(c) there is $t > 0$ such that $\exp t\hat{f}_0 \cdot \xi_0$ is in the relative interior of $R_{\hat{\mathbf{f}}}(\xi_0, t)$,

then there is $\bar{t}$ such that $\exp \bar{t}f_0 \cdot \xi_0$ is in the relative interior of $R_{\mathbf{f}}(\xi_0, \bar{t})$.

The following corollary suggests Theorem 2.2 as a generalization of the "linearization principle."

COROLLARY 2.3. *Let* $M = \mathbb{R}^n$ *and* $f_0(\xi_0) = 0$. *If* $\hat{f}_0$ *is the linear part of* f_0 *at* ξ_0 *and, for* $i = 1, \cdots, m$, $\hat{f}_i$ *is the first significant term in the Taylor approximation of* f_i *at* ξ_0, *then the small time local controllability of* $(\Sigma_{\mathbf{f}}, \mathbb{R}^m)$ *at* ξ_0 *implies the small time local controllability of* $(\Sigma_{\mathbf{f}}, \mathbb{R}^m)$ *at* ξ_0. *Moreover, if* $\mathbf{f}_i$ *is the first significant term of the Taylor approximation of* f_i *at* ξ_0, $i = 0, \cdots, m$, *and if there is a* t *such that* $\xi_0 \in \operatorname{int} R_{\hat{\mathbf{f}}}(\xi_0, t)$, *then there is a* $\bar{t}$ *such that* $\xi_0 \in \operatorname{int} R_{\mathbf{f}}(\xi_0, \bar{t})$.

Proof. It is sufficient to apply Theorem 2.2 and the result in Remark 2.5 to the trivial gradation $w^i = 1$ for each i.

Example 2.2. Let us consider the system given in Example 1.4. If we define on $\mathbb{R}^3$ a graded structure by $\mathbf{w}=(1,2,5)$, we have that the graded order of both f_0 and f_1 is one. The approximating system turns out to be

$$(\Sigma_{\hat{\mathbf{f}}}) \qquad \dot{x}=1+u, \quad \dot{y}=ux, \quad \dot{z}=uy^2.$$

It is easy to see that all the elements in $\mathcal{B}_{\hat{\mathbf{f}}}^{*}$ are zero at ξ_0, so that we can prove that the approximating system is locally controllable along the trajectory $t\mapsto(t,0,0)$. Applying Theorem 2.2 we obtain the local controllability of the original system.

The approximating system depends strongly both on the coordinates $\mathbf{x}$ and on the weights $\mathbf{w}$. Hence we have many choices in the construction of $(\Sigma_{\hat{\mathbf{f}}})$.

In the next section we will define a graded structure linked to the relations at ξ_0 in Lie $\mathbf{f}$.

3. Graded structure adapted to a filtration of a Lie algebra. Let L be a subalgebra of $\mathcal{V}(M)$. An *increasing filtration* of L is a sequence $\mathcal{L}=\{L_i\}_{i\geqq 0}$ of subspaces of L such that

(a) $L_i\subseteq L_{i+1}$,
(b) $[L_i, L_j]\subseteq L_{i+j}$,
(c) $\cup_{i\geqq 0} L_i = L$.

Setting

$$A_i=\operatorname{span}\{Z_1\cdot\cdots\cdot Z_s\colon Z_k\in L_{j_k}, j_1+\cdots+j_s\leqq i\},$$

we get an increasing filtration $\mathcal{A}=\{A_i\}_{i\geqq 0}$ of the subalgebra A of $\mathcal{D}(M)$ generated by L.

Let $\xi_0\in M$ and $m_j=\dim L_j(\xi_0)$, $j=0,\cdots,p$.

LEMMA 3.1. *For each $\alpha\in\mathcal{F}(M)$ such that*

(a) $d\alpha(\xi_0)\neq 0$,
(b) $Z\cdot\alpha(\xi_0)=0$, *for all* $Z\in L_p$,

there is a neighbourhood U of ξ_0 and $\hat{\alpha}\in\mathcal{F}(U)$ such that

(c) $d\hat{\alpha}(\xi_0)=d\alpha(\xi_0)$,
(d) $S\cdot\hat{\alpha}(\xi_0)=0$, *for all* $S\in A_p$.

Proof. Let $\{g_1,\cdots,g_n\}\subset\mathcal{V}(M)$ be a basis of TM at ξ_0 such that

(i) $\{g_1,\cdots,g_{m_i}\}$ is a basis of L_i at ξ_0, $i\leqq p$,
(ii) $g_j\cdot\alpha(\xi_0)=0$, $j=1,\cdots,n-1$,
(iii) $g_n\cdot\alpha(\xi_0)=1$.

Let the chart $\mathbf{x}=(x^1,\cdots,x^n)$ be defined as the local inverse of the map

$$(x^1,\cdots,x^n)\mapsto\exp x^n g_n\cdot\cdots\cdot\exp x^1 g_1\cdot\xi_0.$$

The function $\hat{\alpha}\equiv x^n$ obviously satisfies (c). Let $S=Z_s\cdots Z_1\in A_p$; with $Z_k\in L_{i_k}$ and $i_1+\cdots+i_s\leqq p$, we will prove (d) by induction on s. Parts (a) and (c) imply that (d) is satisfied for $s=1$.

Since $Z_s\in L_{i_s}$, $Z_s(\xi_0)=\sum_{j=1}^{m_{i_s}} a_j g_j(\xi_0)$ for some a_j and

$$S\cdot\hat{\alpha}(\xi_0)=\sum_{j=1}^{m_r} a_j(Z_{s-1}\cdot g_j\cdot Z_{s-2}\cdots Z_1+[g_j,Z_{s-1}]\cdot Z_{s-2}\cdots Z_1)\cdot\hat{\alpha}(\xi_0)$$

$$=\text{(by the induction hypothesis)}$$

$$=\sum_{j=1}^{m_r} a_j Z_{s-1}\cdot g_j\cdots Z_1\cdot\hat{\alpha}(\xi_0).$$

Iterating the procedure we can get S as a linear combination of elements of the type

$$g_{j_s} \cdots g_{j_1} \cdot \hat{\alpha}(\xi_0),$$

with $n > m_p \geqq j_s \geqq \cdots \geqq j_1 \geqq 1$. Therefore we get

$$S \cdot \hat{\alpha}(\xi_0) = S \cdot x^n(\xi_0) = 0. \qquad \square$$

Remark 3.1. The function $\hat{\alpha}$ is not unique. In [22] a way of obtaining $\hat{\alpha}$ by means of an algorithm is described. The algorithm can be applied if for each $Z \in L_0$, $Z(\xi_0) = 0$.

COROLLARY 3.1. *There exists a chart* $\mathbf{x} = (x^1, \cdots, x^n)$ *at* ξ_0 *such that if* $j = 0, \cdots, p$, *then*

(a) $L_j(\xi_0) = \text{span}\{\partial/\partial x^1(\xi_0), \cdots, \partial/\partial x^{m_j}(\xi_0)\}$,

(b) $S \cdot x^k(\xi_0) = 0$, *for all* $S \in A_j$ *and* $k > m_j$.

Proof. Starting from any chart at ξ_0 we can get a chart with the property (a) by means of a linear change of coordinates. Applying Lemma 2.1 to each function of this chart we get the statement. $\square$

DEFINITION 3.1. A chart at ξ_0 with the properties (a) and (b) of Corollary 3.1 is called a chart *adapted to* $\mathscr{L}$ *at* ξ_0 *up to weight* p.

Remark 3.2. Obviously, also the adapted chart is not unique. Applying the algorithm described in [22] we can get an adapted chart by means of a polynomial change of coordinates whose inverse is also polynomial. Moreover, using a proof very similar to the one in Lemma 3.1, an adapted chart can be obtained as a local inverse of the map $(x^1, \cdots, x^n) \mapsto \exp x^n g_n \cdot \cdots \cdot \exp x^1 g_1 \cdot \xi_0$ where $\{g_1, \cdots, g_n\}$ is a basis of TM at ξ_0 with the property (i) stated in the proof of Lemma 3.1.

DEFINITION 3.2. Let $\mathscr{L}$ be such that $L_0(\xi_0) = \{0\}$ and let $(\mathbf{x}, U)$ be a chart adapted to $\mathscr{L}$ at ξ_0 up to weight p such that $\mathbf{x}(U)$ is a ball centred at zero. Let $\mathbf{w}$ be defined by

$$w^j = i \text{ for each } j \in \{m_{i-1}+1, \cdots, m_i\}, \quad i = 1, \cdots, p, \quad w^j = p+1 \text{ for each } j > m_p.$$

The local graded structure $(\mathbf{x}, U, \mathbf{w})$ will be called a graded structure induced by $\mathscr{L}$ at ξ_0 up to weight p.

LEMMA 3.2. *If* $(\mathbf{x}, U, \mathbf{w})$ *is a graded structure induced by* $\mathscr{L}$ *at* ξ_0 *up to weight* p *then*

(a) $\max\{w^1, \cdots, w^n\} \leqq p+1$,

(b) $L_j(\xi_0) = \{h(\xi_0) : h \in \mathscr{V}(U), \mathscr{O}(h) \leqq j\}, j = 1, \cdots, p$,

(c) $S \cdot x^k(\xi_0) = 0$ *for all* $S \in A_{w^k - 1}$.

Proof. The proof is obvious by the definitions. $\square$

Remark 3.3. To define a graded structure induced by $\mathscr{L}$ we need that $L_0(\xi_0) = \{0\}$ because otherwise we could have a coordinate with null weight.

THEOREM 3.1. *If* $f \in L_i$, *then* $\mathscr{O}(f) \leqq i$, *where* $\mathscr{O}$ *is the graded order associated to the graded structure* $(\mathbf{x}, U, \mathbf{w})$.

Proof. Since for all $f \in \mathscr{V}(U)$, $\mathscr{O}(f) \leqq p+1$, we need to prove the theorem for $i = 0, \cdots, p$. Since $\mathscr{O}(f) \leqq i$ if and only if $\mathscr{O}(f^k) = \mathscr{O}(f \cdot x^k) \geqq w^k - i$, $k = 1, \cdots, n$, therefore it is sufficient to prove that for each $k = 1, \cdots, n$ and each $h_1, \cdots, h_s \in \mathscr{V}(U)$ such that $\mathscr{O}(h_1 \cdots h_2) \leqq w^k - i - 1$, we have

$$h_1 \cdots h_s \cdot f \cdot x^k(\xi_0) = 0. \tag{3.1}$$

Let us prove (3.1) by induction on s. Let $s = 1$ and $\mathscr{O}(h_1) \leq w^k - i - 1$. By Lemma 3.2(a) there is $g_1 \in L_{w^k - 1}$ such that $h_1(\xi_0) = g_1(\xi_0)$ so that $h_1 \cdot f \cdot x^k(\xi_0) = g_1 \cdot f \cdot x^k(\xi_0)$ and we get (3.1) by Lemma 3.2(b). Let (3.1) hold for a given s. If $\mathscr{O}(h_1 \cdots h_{s+1}) \leqq w^k - i - 1$, we get

$$h_1 \cdots h_{s+1} \cdot f \cdot x^k(\xi_0) = g_1 \cdots g_{s+1} \cdot f \cdot x^k(\xi_0) + \text{terms of the type } \hat{h}_1 \cdots \hat{h}_s \cdot f \cdot x^k(\xi_0)$$

where $g_1 \cdots g_{s+1}$ belongs to $A_{w^k - 1}$ and $\mathscr{O}(\hat{h}_1 \cdots \hat{h}_s) \leqq w^k - i - 1$. Therefore we get (3.1) by Lemma 3.2(b) and the induction hypothesis. $\square$

Let us now consider the case when the filtration is induced by a set of weights. Let $\mathbf{f}$, $\mathbf{l}$, and $\mathbf{X}$ be as in § 1. If $\mathbf{l}$ is a set of *nonnegative* weights it defines a filtration $\mathscr{L}$ on Lie $\mathbf{f}$ by setting

$$L_i = \text{span}\,\{\Lambda_{\mathbf{f}}\colon \Lambda \in \text{Lie}\, X,\ \|\Lambda\|_{\mathbf{l}} \leqq i\}.$$

$\mathscr{L} = \{L_i\}_{i \geqq 0}$ is a filtration on Lie $\mathbf{f}$ and if $L_0(\xi_0) = \{0\}$ then it induces a graded structure at ξ_0. Theorem 3.1 implies that $\mathscr{O}(f_i) \leqq l_i$, $i = 0, \cdots, m$, therefore we can define the family $\hat{\mathbf{f}}$ as in § 2. Recalling that $\hat{f}_i \in \mathscr{V}(l_i)$, it is easy to prove the following result that summarizes the main properties of the approximating systems induced by a set of weights.

PROPOSITION 3.1. (i) $l_0, \cdots, l_m \geqq 1 \Rightarrow \text{Lie}\,\hat{\mathbf{f}}$ *is nilpotent.*

(ii) $l_1, \cdots, l_m \geqq 1 \Rightarrow \mathscr{S}_{\hat{\mathbf{f}}}$ *is nilpotent.*

(iii) *For each bracket* Λ *subject to* $\|\Lambda\|_{\mathbf{l}} = i$, $(\Lambda_{\hat{\mathbf{f}}} - \Lambda_{\hat{\mathbf{f}}})(\xi_0) \in L_{i-1}(\xi_0)$ *and*

$$\Lambda_{\mathbf{f}}(\xi_0) \in L_{i-1}(\xi_0) \Rightarrow \Lambda_{\hat{\mathbf{f}}}(\xi_0) = 0.$$

(iv) *If* $L_p(\xi_0) = \text{Lie}\,\mathbf{f}(\xi_0)$, *then* $\text{Lie}\,\hat{\mathbf{f}}(\xi_0) = \text{Lie}\,\mathbf{f}(\xi_0)$.

Remark 3.4. The properties described in Proposition 3.1 suggest that the system $(\Sigma_{\hat{\mathbf{f}}})$ is much easier to handle than $(\Sigma_{\mathbf{f}})$. Therefore we can try to prove the local controllability of $(\Sigma_{\hat{\mathbf{f}}})$ and then apply Theorem 2.2. The problem that might arise is that the approximation induced by $\mathbf{l}$ does not guarantee that $\dim \mathscr{S}_{\hat{\mathbf{f}}}(\xi_0) = \dim \mathscr{S}_{\mathbf{f}}(\xi_0)$. As it will be seen in the next section, this problem can be avoided by introducing an auxiliary system. However, note that if the reference trajectory is stationary, i.e., $f_0(\xi_0) = 0$, then

$$\text{Lie}\,\mathbf{f}(\xi_0) = \mathscr{S}_{\mathbf{f}}(\xi_0) = \text{Lie}\,\hat{\mathbf{f}}(\xi_0) = \mathscr{S}_{\hat{\mathbf{f}}}(\xi_0),$$

see [21].

We end this section by summarizing the theory with an example showing how the local controllability properties are more evident in an adapted chart.

The following system is exhibited in [3] as an example of a small time locally controllable bilinear system for which the property is destroyed by a second-order perturbation.

Example 3.1. Let $M = \mathbb{R}^2$, $\xi_0 = (0, 0)$, $\Omega = \{\omega \in \mathbb{R}\colon |\omega| \leqq 1\}$ and

$$f_0 = x\frac{\partial}{\partial x} + 2y\frac{\partial}{\partial y}, \qquad f_1 = \frac{\partial}{\partial x} + (x+y)\frac{\partial}{\partial y}.$$

The significant Lie brackets are

$$[f_1, f_0] = \text{ad}^2_{f_0} f_1 = \frac{\partial}{\partial x} + x\frac{\partial}{\partial y}, \quad \text{ad}^2_{f_1} f_0 = -x\frac{\partial}{\partial y} = (-1)^k\,\text{ad}^k_{f_0}\,\text{ad}^2_{f_1} f_0, \quad \text{ad}^3_{f_1} f_0 = (-1+x)\frac{\partial}{\partial y}.$$

It is not difficult to see that Theorem 1.1 (or the result in [24]) can be applied with $\mathbf{l} = (1, 1)$. Perturbing f_0 with the term $x^2(\partial/\partial y)$ we obtain a new system that is not small time locally controllable (see [25], [22]). The reason will be evident when we rewrite the system in the adapted chart induced by $\mathbf{l}$. Set $g_1 = f_1$ and $g_2 = -\text{ad}^3_{f_1} f_0 = (1-x)(\partial/\partial y)$. The adapted chart suggested by Corollary 3.1 is given by the local inverse of the map

$$(x(\alpha, \beta), y(\alpha, \beta)) = \exp \beta g_2 \cdot \exp \alpha g_1 \cdot \xi_0 = (\alpha, (1-\alpha)\beta + e^{\alpha} - \alpha - 1)$$

and the graded structure is defined by $\mathbf{w} = (1, 4)$.

In the new coordinates the vector fields become

$$f_0 = \alpha\frac{\partial}{\partial\alpha} + \frac{\beta(2-\alpha)+2(e^\alpha-\alpha-1)+(1-e^\alpha)}{(1-\alpha)}\frac{\partial}{\partial\beta}$$

$$= \alpha\frac{\partial}{\partial\alpha} + \frac{\beta(2-\alpha)+\alpha^2+\alpha^3/3+\cdots-\alpha^2-\alpha^3/6-\cdots}{(1-\alpha)}\frac{\partial}{\partial\beta}$$

$$= \frac{\alpha^3}{6}\frac{\partial}{\partial\beta} + \text{a vector field of order 0.}$$

$$f_1 = \frac{\partial}{\partial\alpha} + \frac{\beta(2-\alpha)}{(1-\alpha)}\frac{\partial}{\partial\beta} = \frac{\partial}{\partial\alpha} + \text{a vector field of order 0.}$$

Therefore the approximating system is given by

$$(\Sigma_{\hat{\mathbf{f}}}) \qquad \dot{\alpha} = u, \qquad \dot{\beta} = \frac{\alpha^3}{6}.$$

The perturbating vector field $x^2(\partial/\partial y) = \alpha^2/(1-\alpha)\,\partial/\partial\beta$ has graded order two, hence it affects the graded order of f_0. On the other hand in the new coordinates the perturbed system is given by

$$\dot{\alpha} = \alpha + u,$$

$$\dot{\beta} = \frac{\beta(2-\alpha)+\alpha^2+\alpha^3/6+\cdots}{(1-\alpha)} + u\frac{\beta(2-\alpha)}{(1-\alpha)}.$$

Therefore it is evident that the perturbation does destroy the small time local controllability of the original system.

Note that applying the algorithm described in [22] we obtain for the adapted chart induced by $\mathbf{l}$

$$\gamma = x, \qquad \delta = y - \frac{1}{2}x^2 - \frac{1}{3!}x^3.$$

This chart is global whereas the one above is local. With this choice of the adapted chart the approximating system becomes

$$(\Sigma_{\hat{\mathbf{f}}}) \qquad \dot{\gamma} = u, \qquad \dot{\delta} = \frac{\gamma^3}{6} + u\frac{\gamma^3}{3!}$$

and the perturbed system becomes

$$\dot{\gamma} = \gamma + u, \qquad \dot{\delta} = 2\delta + \gamma^2 + \frac{\gamma^3}{6} + u\frac{\gamma^3}{6}.$$

4. Proof of the main theorem. We start this section by reformulating the problem by means of the "pull-back" system of $(\Sigma_{\mathbf{f}})$. The controllability properties of $(\Sigma_{\mathbf{f}}, \Omega)$ can be investigated much better by means of this new system. In fact it describes the behaviour of the trajectories of $(\Sigma_{\mathbf{f}})$ relative to the reference trajectory and it clarifies the role of the submanifold $N_{\mathbf{f}}^0(\xi_0)$ with respect to local controllability.

By the properties of differential equations it follows that there is a neighbourhood U of ξ_0 and a neighbourhood I of zero in $\mathbb{R}$ containing J such that $\exp tf_0 \cdot \xi$ is defined

for each t in I and each ξ in U. Therefore we can define for any $f \in \mathscr{V}(M)$ the time-dependent vector field

$$f^*: I \times U \to TU \subset TM, \qquad f^*(t, \xi) \equiv f^*(t)(\xi) = \exp(-tf_0)_* \cdot f(\exp tf_0 \cdot \xi).$$

It is easy to see that whatever are $f, g \in \mathscr{V}(M)$

$$[f^*, g^*] = [f, g]^*, \tag{4.1}$$

$$\frac{\partial f^*}{\partial t} = (\mathrm{ad}_{f_0} f)^*. \tag{4.2}$$

Moreover, it is not difficult to prove the following property:

$$\forall \Lambda \in \mathscr{S} \quad \Lambda_{\mathbf{f}^*}(t)(\xi) = \exp(-tf_0)_* \Lambda_{\mathbf{f}}(\exp tf_0 \cdot \xi). \tag{4.3}$$

If u belongs to a sufficiently small neighbourhood of zero in $L^1([0, T], \Omega)$, the map

$$t \mapsto y(t, \xi_0, u) \equiv \exp(-tf_0)(s_{\mathbf{f}}(t, \xi_0, u))$$

is defined and it satisfies the time-dependent control system on U

$$\dot{y}(t) = \sum_{i=1}^{m} u_i(t) f_i^*(t, y(t)), \qquad y(0) = \xi_0.$$

The time-dependent vector field f_i^* may be viewed as a C^∞ vector field on $M^* = I \times U$. If we set $f^* = \{\partial/\partial t, f_1^*, \cdots, f_m^*\}$, the pullback system of the system $(\Sigma_{\mathbf{f}})$ is the system $(\Sigma_{\mathbf{f}^*})$ on M^*, given by

$$(\Sigma_{\mathbf{f}^*}) \qquad \dot{x} = \frac{\partial}{\partial t} + \sum_{i=1}^{m} u_i f_i^*(x).$$

The solutions of the system $(\Sigma_{\mathbf{f}})$ and those of the system $(\Sigma_{\mathbf{f}^*})$ are linked by the relation

$$s_{\mathbf{f}^*}(t, (\tau, \xi_0), u) = (t + \tau, \exp(-(t+\tau)f_0) \cdot s_{\mathbf{f}}(t, \exp \tau f_0 \cdot \xi_0, u)). \tag{4.4}$$

In what follows we will identify U with $\{0\} \times U$, so that $\xi_0 \equiv (0, \xi_0)$, $N^0_{\mathbf{f}^*}(\xi_0)$ is contained in $N^0_{\mathbf{f}}(\xi_0)$ and $N_{\mathbf{f}^*}(\xi_0, t) = \{t\} \times N^0_{\mathbf{f}^*}(\xi_0)$.

From (4.4) it follows that $(\Sigma_{\mathbf{f}}, \Omega)$ is weakly locally controllable along the trajectory $\hat{x}$ if $(\Sigma_{\mathbf{f}^*}, \Omega)$ is weakly locally controllable along the trajectory $t \mapsto (t, \xi_0)$.

LEMMA 4.1. *$(\Sigma_{\mathbf{f}}, H_\rho)$ satisfies the assumptions of Theorem* 1.1 *if and only if $(\Sigma_{\mathbf{f}^*}, H_\rho)$ does.*

Proof. The statement is obvious taking into account that (4.3) implies that $\chi_{\mathbf{f}^*}(\xi_0) = \chi_{\mathbf{f}}(\xi_0)$ for each $\chi \in \mathscr{S}$. □

LEMMA 4.2. *If* (b) *of Theorem* 1.1 *is satisfied for a set of admissible nonnegative weights* $\mathbf{l}$, *then it is satisfied for a set of admissible positive weights.*

Proof. Let $J = \{i \in \{0, \cdots, m\}: l_i = 0\}$. If $J = \varnothing$, nothing has to be proved. Let $J \neq \varnothing$; for each $\Lambda \in \mathscr{S}$, let $p(\Lambda) = \sum_{i \in J} |\Lambda|_i$. Choose $\Lambda_1, \cdots, \Lambda_s \in \mathscr{S}$ in such a way that $\Lambda_{1\mathbf{f}}(\xi_0), \cdots, \Lambda_{mi\mathbf{f}}(\xi_0)$ is a basis for $V_i^{\mathbf{l}}(\xi_0)$ and $\Lambda_{1\mathbf{f}}(\xi_0), \cdots, \Lambda_{s\mathbf{f}}(\xi_0)$ is a basis for $\mathscr{S}_{\mathbf{f}}(\xi_0)$.

Let $h > \max\{p(\Lambda_i),\ i = 1, \cdots, s\}$ and set $l_i' = 1$ if $i \in J$ and $l_i' = hl_i$ otherwise. If $\mathbf{l}$ is admissible for H_ρ then $\mathbf{l}'$ is admissible for H_ρ. Moreover if the hypothesis (b) of Theorem 1.1 is fulfilled for the set $\mathbf{l}$, it is fulfilled for the set $\mathbf{l}'$. In fact let $\Theta \in \mathscr{B}^*_{\mathbf{l}'} = \mathscr{B}^*_{\mathbf{l}}$. The hypothesis (b) of Theorem 1.1 implies

$$\Theta_{\mathbf{f}}(\xi_0) = \sum b_j \Lambda_{j\mathbf{f}}(\xi_0) \quad \text{with } \|\Lambda_j\|_{\mathbf{l}} < \|\Theta\|_{\mathbf{l}}.$$

If $p(\Theta) \geqq p(\Lambda_j)$, then

$$\|\Theta\|_{\mathbf{l}'} = h\|\Theta\|_{\mathbf{l}} + p(\Theta) > h\|\Lambda_j\|_{\mathbf{l}} + p(\Lambda_j) = \|\Lambda_j\|_{\mathbf{l}'}$$

and if $p(\Theta) < p(\Lambda_j)$, then

$$\|\Theta\|_{\mathbf{I}'} = h\|\Lambda_j\|_{\mathbf{I}} + h(\|\Theta\|_{\mathbf{I}} - \|\Lambda_j\|_{\mathbf{I}}) + p(\Theta) \geqq h\|\Lambda_j\|_{\mathbf{I}} + h + p(\Theta) > \|\Lambda_j\|_{\mathbf{I}'}. \qquad \square$$

Lemmas 4.1 and 4.2 imply that we can prove the theorem for the system $(\Sigma_{\mathbf{f}^*}, H_\rho)$ under the hypothesis that $\mathbf{I}$ is a set of admissible *positive* weights. The sketch of the proof is the following:

Step 1. We define a graded structure $(\mathbf{x}, U, \mathbf{w})$ at ξ_0 by means of the filtration $\mathcal{N} = \{N_i\}_{i \geqq 0}$ of $\mathcal{S}_{\mathbf{f}}$, where $N_i = \operatorname{span}\{\Lambda_{\mathbf{f}}\colon \Lambda \in \mathcal{S},\ \|\Lambda\|_{\mathbf{I}} \leqq i\}$.

We extend $(\mathbf{x}, U, \mathbf{w})$ to a graded structure $(\mathbf{x}^*, M^*, \mathbf{w}^*)$ at ξ_0.

Step 2. We prove that the approximating family $\hat{\mathbf{f}}^* = \{\partial/\partial t, \hat{f}_1^*, \cdots, \hat{f}_m^*\}$ has the property

$$\mathcal{S}_{\hat{\mathbf{f}}^*}(\xi_0) = \mathcal{S}_{\mathbf{f}^*}(\xi_0).$$

Step 3. We prove that $(\Sigma_{\hat{\mathbf{f}}^*}, H_\rho)$ is weakly locally controllable along $t \mapsto (t, \xi_0)$, so that Theorem 1.1 follows by Theorem 2.2.

Let $(\mathbf{x}, U, \mathbf{w})$ be a graded structure induced by $\mathcal{N}$ at ξ_0 up to weight $p = \min\{i\colon N_i(\xi_0) = \mathcal{S}_{\mathbf{f}}(\xi_0)\}$. Without loss of generality we can suppose that the pull-back system $\Sigma_{\mathbf{f}^*}$ is defined on $M^* = I \times U$. Denoting by $x^0\colon M^* \to \mathbb{R}$ the first canonical projection, we set $\mathbf{x}^* = (x^0, x)$ and $\mathbf{w}^* = (l_0, \mathbf{w})$. In the chart $\mathbf{x}^*$ the vector field $\partial/\partial t$ is $\partial/\partial x^0$. $(\mathbf{x}^*, M^*, \mathbf{w}^*)$ defines a local graded structure on M^* at ξ_0.

LEMMA 4.3. *For each $f \in \mathcal{V}(M)$, $\mathcal{O}(f^*) = \mathcal{O}(f)$, where $\mathcal{O}$ denotes both the graded orders induced by the above graded structures on M and M^*.*

Proof. The asymptotic expansion of f^* with respect to t is given by

$$f^*(t, \xi) = f(\xi) + \sum_{k>0} \frac{t^k}{k!} \operatorname{ad}_{f_0}^k f(\xi).$$

By the property (P1) of the graded structure, $\mathcal{O}(\operatorname{ad}_{f_0}^k f) \leqq k l_0 + \mathcal{O}(f)$, hence $\mathcal{O}((x^0)^k \operatorname{ad}_{f_0}^k f) \leqq \mathcal{O}(f)$ for each $k > 0$ so that the statement is proved. $\square$

LEMMA 4.4. *$(\mathbf{x}^*, M^*, \mathbf{w}^*)$ is a local graded structure induced by $\mathcal{L}^* = \{L_i^*\}_{i \geqq 0}$ at ξ_0 up to weight p^*, where*

$$L_i^* = \operatorname{span}\{\Lambda_{\mathbf{f}^*}\colon \Lambda \in \operatorname{Lie} \mathbf{X},\ \|\Lambda\|_{\mathbf{I}} \leqq i\} \quad \textit{and} \quad p^* = \max\{l_0, p\}.$$

Proof. Let $N_i^* = \operatorname{span}\{\Lambda_{\mathbf{f}^*}\colon \Lambda \in \mathcal{S},\ \|\Lambda\|_{\mathbf{I}} \leqq i\}$. By property (4.3) it follows that $N_j(\xi_0) = N_j^*(\xi_0)$. Hence if $j < l_0$, then

$$L_j^*(\xi_0) = N_j^*(\xi_0) = \operatorname{span}\left\{\frac{\partial}{\partial x^1}, \cdots, \frac{\partial}{\partial x^{m_j}}\right\},$$

and if $j \geqq l_0$, then

$$L_j^*(\xi_0) = \operatorname{span}\left\{\frac{\partial}{\partial t}\right\} + N_j^*(\xi_0) = \operatorname{span}\left\{\frac{\partial}{\partial x^0}, \cdots, \frac{\partial}{\partial x^{m_j}}\right\}.$$

For each bracket $\Lambda \in \mathcal{S}$, $\Lambda_{\mathbf{f}^*|U} = \Lambda_{\mathbf{f}|U}$, therefore if $\Lambda_1, \cdots, \Lambda_s \in \mathcal{S}$, then

$$\Lambda_{1\mathbf{f}^*} \cdots \Lambda_{s\mathbf{f}^*} \cdot x^k(\xi_0) = \begin{cases} 0 & \text{if } k = 0, \\ \Lambda_{1\mathbf{f}} \cdots \Lambda_{s\mathbf{f}} \cdot x^k(\xi_0) & \text{if } k \neq 0. \end{cases} \tag{4.5}$$

Moreover, for any $\Lambda \in \operatorname{Lie} \mathbf{X}$

$$\frac{\partial}{\partial t} \cdot \Lambda_{\mathbf{f}^*} \cdot x^k \equiv \Lambda_{\mathbf{f}^*} \cdot \frac{\partial}{\partial t} \cdot x^k + [X_0, \Lambda]_{\mathbf{f}^*} \cdot x^k \equiv [X_0, \Lambda]_{\mathbf{f}^*} \cdot x^k,$$

so that (4.5) proves the statement. $\square$

COROLLARY 4.1. *Let $\hat{f}_i^*$ be the graded approximation of weight l_i of f_i^* and let $\hat{\mathbf{f}}^* = \{\partial/\partial t, \hat{f}_1^*, \cdots, \hat{f}_m^*\}$; then*

(a) $\mathscr{S}_{\hat{\mathbf{f}}^*}(\xi_0) = \mathscr{S}_{\mathbf{f}^*}(\xi_0) = \mathscr{S}_{\mathbf{f}}(\xi_0)$.

(b) $\mathscr{S}_{\hat{\mathbf{f}}^*}$ *is nilpotent of step* p^*.

(c) *If* (b) *of Theorem* 1.1 *holds, then for all* $\chi \in B_1^*$, $\chi_{\hat{\mathbf{f}}^*}(\xi_0) = 0$.

Proof. Part (a) follows by $N_j^*(\xi_0) = N_j(\xi_0)$. Part (b) follows by Proposition 3.1(ii). Part (c) follows by Proposition 3.1(iii). □

THEOREM 4.1. *Let the assumptions of Theorem* 1.1 *hold for a set of admissible positive weights. Then $(\Sigma_{\hat{\mathbf{f}}^*}, H_\rho)$ is weakly locally controllable along the trajectory $t \mapsto (t, \xi_0)$.*

Proof. First of all we note that $(\Sigma_{\hat{\mathbf{f}}^*})$ is a cascade of integrators (see Remark 2.1) so that it can be thought of as the restriction on U of an analytic system (Σ_ϕ) on $\mathbb{R}^n$. The trajectories of $(\Sigma_{\hat{\mathbf{f}}^*})$ coincide with the trajectories of (Σ_ϕ) that remain in U. Moreover, the vector fields of ϕ are complete and generate a nilpotent Lie algebra, so that Lemma 1.1 applies.

Let K be the natural number given by Lemma 1.1 and let $H, \beta = (\beta_1, \cdots, \beta_{r+1})$ be as in property (c) of the Appendix. Set

$$\sigma = \max\{|(\beta_i)_j| : i = 1, \cdots, r+1, j = 1, \cdots, m\}$$

and choose $\tau \in \mathbb{R}^+$ such that $K\tau < H$ and

$$s_\phi(t, \xi_0, u) \in U \quad \forall t \in [0, K\tau] \quad \text{and} \quad \forall u \in L^1([0, K\tau], H_\sigma). \tag{4.6}$$

Again by property (c) in the Appendix we can choose $\bar{\mathbf{b}} \in B_T$, $T < \tau$, such that the map $g_{|B_T}$ has rank r at $\bar{\mathbf{b}}$.

Applying Lemma 1.1 to the pair $(\beta, \bar{\mathbf{b}})$ we get a $\chi \in B_S^1$ and a pair $(\gamma, \mathbf{c})$ such that

$$s_\phi(LT, \xi_0, (\beta\gamma, \bar{\mathbf{b}}\mathbf{c})) = s_{\hat{\mathbf{f}}^*}(LT, \xi_0, (\beta\gamma, \bar{\mathbf{b}}\mathbf{c})) = \exp\left(LT\frac{\partial}{\partial t} + \chi_{\hat{\mathbf{f}}^*}\right) \cdot \xi_0.$$

Therefore there is $\Gamma \in B_1^*$ such that

$$\exp\left(-LT\frac{\partial}{\partial t}\right) \cdot s_{\hat{\mathbf{f}}^*}(LT, \xi_0, (\beta\gamma, \bar{\mathbf{b}}\mathbf{c})) = \exp \Gamma_{\hat{\mathbf{f}}^*} \cdot \xi_0.$$

Let us consider the map $G : B_{LT} \to U$ given by

$$G : \mathbf{t} \mapsto \exp\left(-LT\frac{\partial}{\partial t}\right) \cdot s_{\hat{\mathbf{f}}^*}((L-1)T, s_{\hat{\mathbf{f}}^*}(T, \xi_0, (\beta, \mathbf{t})), (\gamma, \mathbf{c})).$$

Since Corollary 4.1(c) implies $\Gamma_{\hat{\mathbf{f}}^*}(\xi_0) = 0$, we have

$$G(\bar{\mathbf{b}}) = \exp \Gamma_{\hat{\mathbf{f}}^*} \cdot \xi_0 = \xi_0.$$

Moreover, G has rank r at $\bar{\mathbf{b}}$. Hence $\exp(LT(\partial/\partial t)) \cdot G(B_{LT}) \subset R_{\mathbf{f}^*}(\xi_0, LT)$ is a neighbourhood of (LT, ξ_0) in $N_{\mathbf{f}^*}(\xi_0, LT)$. Therefore the reference trajectory is in the relative interior of the reachable set at time LT. Theorem 2.1(b) and $l_0 > 0$ imply that it is in the relative interior at any time less than LT. □

Appendix. Let $\mathbf{f} = \{f_0, f_1, \cdots, f_m\}$ be a family of analytic vector fields on $\mathbb{R}^n$ and let Ω be as in § 1. To each pair $(\alpha, \mathbf{a})$, $\alpha \in \Omega^k$, $\mathbf{a} \in (\mathbb{R}_+)^k$, $k \in \mathbb{N}$, we associate the piecewise constant control defined in $[0, \sum_{i=1}^k a_i]$ which assumes the value α_i in the interval $[\sum_{j=1}^{i-1} a_j, \sum_{j=1}^{i} a_j]$. We will use the same notation for both the pair and the control. Hence

$$s_{\mathbf{f}}\left(\sum_{i=1}^k a_i, y, (\boldsymbol{\alpha}, \mathbf{a})\right) = \exp a_k\left(f_0 + \sum_{i=1}^m (\alpha_k)_i f_i\right) \cdots \exp a_1\left(f_0 + \sum_{i=1}^m (\alpha_1)_i f_i\right) \cdot y.$$

A pair will be called a k-pair if it belongs to $\Omega^k \times (\mathbb{R}_+)^k$. Given a k-pair $(\boldsymbol{\alpha}, \mathbf{a})$ and an s-pair $(\boldsymbol{\beta}, \mathbf{b})$ we set $(\boldsymbol{\alpha}\boldsymbol{\beta}, \mathbf{ab})$ the $(k+s)$-pair

$$(\alpha_1, \cdots, a_k, \beta_1, \cdots, \beta_s, a_1, \cdots, a_k, b_1, \cdots, b_s).$$

We recall some properties of the reachable sets in the analytic case.

(a) The connected components of $N_{\mathbf{f}}(\xi, t)$ are integral manifolds of $\mathscr{S}_{\mathbf{f}}$;

(b) $\operatorname{int}_{\text{rel}} R_{\mathbf{f}}(\xi, t)$ is dense in $R_{\mathbf{f}}(\xi, t)$;

(c) If $\dim \mathscr{S}_{\mathbf{f}}(\xi) = r$, there exist $H > 0$ and $\beta \in (\Omega)^{r+1}$ such that for all $\mathbf{t} \in (\mathbb{R}_+)^{r+1}$, $\sum_{i=1}^{r+1} t_i < H$, the map

$$g : \mathbf{t} \to s_{\mathbf{f}}\left(\sum_{i=1}^{r+1} t_i, \xi, (\beta, \mathbf{t})\right)$$

is defined and analytic. Moreover, if B_T is the set $\{t \in \mathbb{R}^{r+1}, t_i > 0, \sum_{i+1}^{r+1} t_i = T\}$, for each $t < H$ there exist $T < t$ and $\bar{\mathbf{b}} \in B_T$ such that g restricted to B_T has rank r at $\bar{\mathbf{b}}$.

For properties (a) and (b) we refer to [27]; property (c) can be derived using the arguments in [27] and [15].

LEMMA A. If $\exp \bar{t} f_0 \cdot \xi \in \operatorname{int}_{\text{rel}} R_{\mathbf{f}}(\bar{t}, \xi)$, *then for all* $t > \bar{t}$ *sufficiently close to* $\bar{t}$, *there exist*

(a) *a* k*-pair* $(\boldsymbol{\omega}, \mathbf{t})$, $\sum_{i=1}^{k} t_i = t$,

(b) *a neighbourhood* V *of* $\exp t f_0 \cdot \xi$ *in* $N_{\mathbf{f}}(\xi, t)$ *and an analytic map* $\tau : V \to (\mathbb{R}_+)^k$ *such that*

$$s_{\mathbf{f}}(t, \xi, (\boldsymbol{\omega}, \tau(v))) = v \quad \forall v \in V.$$

Proof. Let $t > \bar{t}$ be such that $y = \exp t f_0 \cdot \xi$ is defined. If $d = t - \bar{t}$, $R_{\mathbf{f}}(y, -d) \cap R_{\mathbf{f}}(\xi, \bar{t}) \neq \varnothing$ so that

$$W = \operatorname{int}_{\text{rel}} R_{\mathbf{f}}(y, -d) \cap \operatorname{int}_{\text{rel}} R_{\mathbf{f}}(\xi, \bar{t}) \neq \varnothing.$$

Let $z = s_{\mathbf{f}}(\bar{t}, \xi, (\boldsymbol{\alpha}, \mathbf{a})) \in W$. There is a neighbourhood V of ξ in $N_{\mathbf{f}}^0(\xi)$ such that for all $y' \in V$ $s_{\mathbf{f}}(\bar{t}, y', (\boldsymbol{\alpha}, \mathbf{a})) \in W$.

Let $\boldsymbol{\beta}$ be as in (c) and let $\rho > 0$ be such that for each pair $(\boldsymbol{\beta}, t)$ with $\sum_{i=1}^{r+1} t_i \leqq \rho$, $\exp(-\sum_{i=1}^{r+1} t_i)(f_0 + \sum_{i=1}^{m} (\alpha_1)_i f_i) \cdot s_{\mathbf{f}}(\sum_{i=1}^{r+1} t_i, \xi, (\boldsymbol{\beta}, \mathbf{t})) \in V'$. By (c) there exists $h < \min(\rho, a_1)$, and $\bar{\mathbf{b}} \in B_h$, such that the map defined in a neighbourhood of $\bar{\mathbf{b}}$ by $\mathbf{b} \mapsto s_{\mathbf{f}}(h, \xi, (\boldsymbol{\beta}, \mathbf{b}))$ has rank r at $\bar{\mathbf{b}}$. By construction $\bar{y} = s_{\mathbf{f}}(\bar{t}, \xi, (\boldsymbol{\beta}\boldsymbol{\alpha}, \bar{\mathbf{b}}(a_1 - h)a_2 \cdots a_s)) \in W$. Hence there exists a pair $(\boldsymbol{\gamma}, \mathbf{c})$ such that $s_{\mathbf{f}}(d, \bar{y}, (\boldsymbol{\gamma}, \mathbf{c})) = \exp t f_0 \cdot \xi$. Let $\boldsymbol{\omega} = (\boldsymbol{\beta}\boldsymbol{\alpha}\boldsymbol{\gamma})$ and $\mathbf{d}(\mathbf{b}) = (\mathbf{b}, (a_1 - h), a_2, \cdots, a_s, \mathbf{c})$.

The map defined in a neighbourhood of $\bar{\mathbf{b}}$ in B_h by $\mathbf{b} \mapsto s_{\mathbf{f}}(t, \xi, (\omega, \mathbf{d}(\mathbf{b})))$ is an analytic map from a neighbourhood of an r-dimensional analytic manifold to an r-dimensional analytic manifold and has rank equal to r at $\bar{\mathbf{b}}$. The statement follows by the inverse mapping theorem. □

REFERENCES

[1] R. M. BIANCHINI AND G. STEFANI, *Local controllability along a reference trajectory*, in Analysis and Optimization of Systems, Lecture Notes in Control and Information Sciences, 83, Springer-Verlag, Berlin, New York, 1986, pp. 342-352.

[2] ———, *Multivalued fields and control system with unbounded controls*, Ricerche Automat., 12 (1981), pp. 33-49.

[3] ———, *Normal local controllability of order one*, Internat. J. Control, 39 (1984), pp. 701-714.

[4] A. BRESSAN, *Local asymptotic approximation of non linear control systems*, Internat. J. Control, 41 (1985), pp. 1331-1336.

[5] ———, *A high order test for optimality of bang-bang controls*, SIAM J. Control Optim., 23 (1985), pp. 38-48.

[6] P. E. CROUCH AND C. I. BYRNES, *Symmetries and local controllability*, in Algebraic and Geometric Methods in Nonlinear Control Theory, Proc. Conference Paris 1985 Math. Appl., Reidel, Boston, MA, 1986, pp. 55-75.

[7] P. E. CROUCH, *Solvable approximations to control systems*, SIAM J. Control Optim., 22 (1984), pp. 40-54.

[8] J. B. GONCALVES, *Sufficient conditions for local controllability with unbounded controls*, SIAM J. Control Optim., 25 (1987), pp. 1371-1378.

[9] R. W. GOODMAN, *Nilpotent Lie Groups: Structure and Applications to Analysis*, Lecture Notes in Mathematics, 562, Springer-Verlag, Berlin, New York, 1976.

[10] H. HERMES, *On local controllability*, SIAM J. Control Optim., 20 (1982), pp. 211-220.

[11] ———, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166-187.

[12] N. KALOUPTSIDIS AND D. L. ELLIOT, *Accessibility properties of smooth nonlinear control system*, AMES Research Centre 1976, Martin and Hermann, eds., Math. Sci. Press, Brookline, MA, 1977, pp. 439-446.

[13] M. KAWSKI, *A new necessary condition for local controllability*, AMS Contemporary Math. Series, Proc. Conference on Differential Geometry, San Antonio, TX, 1986.

[14] A. KRENER, *Local approximation of control systems*, J. Differential Equations, 19 (1975), pp. 125-133.

[15] ———, *A generalization of Chow's theorem and the bang-bang theorem to non-linear control problems*, SIAM J. Control Optim., 12 (1974), pp. 43-52.

[16] S. LANG, *Real Analysis*, Addison-Wesley, Reading, MA, 1973.

[17] A. A. LIVEROWSKII, *Some properties of Bellman's function for linear and symmetric polysystems*, J. Differential Equations, 16 (1980), pp. 255-261.

[18] N. N. PETROV, *Local controllability*, Differentsial'nye Uravneniya, 12 (1976), pp. 2214-2222. (In Russian.)

[19] C. ROCKLAND, *Intrinsic nilpotent approximations to filtered Lie algebras*, Acta Appl. Math., 8 (1987), pp. 285-311.

[20] L. P. ROTHSCHILD AND E. M. STEIN, *Hypoelliptic differential operators and nilpotent groups*, Acta Math., 137 (1976), pp. 247-320.

[21] G. STEFANI, *Polynomial approximations to control systems and local controllability*, in Proc. 25th IEEE Conference on Decision and Control, Ft. Lauderdale, FL, 1985, pp. 33-38.

[22] ———, *On local controllability of a scalar input control system*, in Theory and Applications of Nonlinear Control Systems, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 167-179.

[23] ———, *On the minimum time problem*, in Analysis and Control of Nonlinear Systems, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North-Holland, Amsterdam, 1988, pp. 213-220.

[24] H. J. SUSSMANN, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158-194.

[25] ———, *Lie brackets and local controllability: a sufficient condition for scalar input systems*, SIAM J. Control Optim., 21 (1983), pp. 686-713.

[26] ———, *Orbits of families of vector fields and integrability of systems with singularities*, Trans. Amer. Math. Soc., 180 (1973), pp. 171-188.

[27] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95-116.

[28] C. VARSAN, *On local controllability for non linear control systems*, Rev. Roumaine Math. Pures Appl., 29 (1984), pp. 823-824.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 925-935, July 1990

000

A COMPARISON OF CONSTRAINT QUALIFICATIONS IN INFINITE-DIMENSIONAL CONVEX PROGRAMMING*

M. SEETHARAMA GOWDA† AND MARC TEBOULLE†‡

Abstract. In this paper the relationships between various constraint qualifications for infinite-dimensional convex programs are investigated. Using Robinson's refinement of the duality result of Rockafellar, it is demonstrated that the constraint qualification proposed by Rockafellar provides a systematic mechanism for comparing many constraint qualifications as well as establishing new results in different topological environments.

Key words. constraint qualifications, convex programming, duality theory

AMS(MOS) subject classifications. 49A27, 49A55, 90C25, 90C48

1. Introduction. This paper deals with constraint qualifications for infinite-dimensional convex programs. A constraint qualification is an essential condition needed to establish strong duality results for a pair of optimization problems. The usual constraint qualification is a *Slater-like* condition that requires nonempty interiority of a certain convex set. Unfortunately, this condition often fails for an important class of optimization problems arising in applications, see, e.g., [3].

Recently, many authors have proposed new constraint qualifications for optimization problems in infinite-dimensional vector spaces, see [5]-[7], [2], [15]. Motivated by the recent constraint qualification proposed by Borwein and Wolkowicz [7], in this paper we investigate the relationships between various constraint qualifications. By studying cores and interiors of convex sets, we show that many of the constraint qualifications are equivalent or can be derived from the constraint qualification proposed by Rockafellar [14]. Furthermore, we show that the Rockafellar constraint qualification provides a natural mechanism for establishing new constraint qualifications in various topological environments.

The paper is organized as follows. In § 2 we recall the fundamental constraint qualification proposed by Rockafellar, denoted by (R), and state Robinson's refinement of a result of Rockafellar. In § 3 we demonstrate that condition (R) is instrumental in constructing various constraint qualifications and that many seemingly unrelated constraint qualifications are in fact related to (R). We also derive new results in the general setting of Baire spaces and provide examples.

2. A fundamental constraint qualification. Let X and Y be real locally convex topological vector spaces and $A: X \to Y$ be a continuous linear operator. Let $f: X \to (-\infty, +\infty]$ and $g: Y \to (-\infty, +\infty]$ be proper, lower semicontinuous convex functions. Consider the primal problem:

$$\text{(P)} \qquad \inf_{x \in X} \{f(x) + g(Ax)\}.$$

The *Fenchel-Rockafellar duality theory*, see Rockafellar [14], associates with (P) the dual problem:

$$\text{(D)} \qquad \sup_{y \in Y^*} \{-g^*(y) - f^*(-A^*y)\}$$

* Received by the editors January 23, 1989; accepted for publication (in revised form) September 18, 1989.

† Department of Mathematics and Statistics, University of Maryland, Baltimore County Campus, Baltimore, Maryland 21228.

‡ This research was supported by Air Force Office of Scientific Research grant 0218-88 and National Science Foundation grant ECS-8802239.

where $A^*: Y^* \to X^*$ is the adjoint of A and X^*, Y^* are the dual spaces of X and Y, respectively. We recall that for a given function $\phi: X \to (-\infty, +\infty]$, the domain is

$$\operatorname{dom} \phi := \{x \in X: \phi(x) < \infty\}$$

and the *conjugate function* is

$$\phi^*(x^*) := \sup_{x \in \operatorname{dom} \phi} \{\langle x^*, x\rangle - \phi(x)\}, \qquad x^* \in X^*.$$

The main issue regarding the pair of problems (P), (D) is the lack of duality gap, i.e., the proof of the strong duality relation

$$\inf_{x \in X} \{f(x) + g(Ax)\} = \max_{y^* \in Y^*} \{-g^*(y^*) - f^*(-A^*y^*)\} \tag{2.1}$$

which we write, for convenience, as

$$\inf (\mathrm{P})^1 = \max (\mathrm{D}).$$

This can be obtained provided a certain constraint qualification (CQ for short) is satisfied. One of the most popular (CQ) is the so-called *Slater* condition: (see, e.g., [1])

$$0 \in \operatorname{int} (\operatorname{dom} g - A \operatorname{dom} f)). \tag{S}$$

THEOREM 2.1. *Suppose that* (S) *holds. Then* $\inf (\mathrm{P}) = \max (\mathrm{D})$.

Unfortunately, in many important applications the Slater condition fails.

A more general constraint qualification was suggested by Rockafellar [14]. Before stating the condition, we recall the definition of the *core* of a set. For a set $C \subset X$, the *core* of C is defined by

$$\operatorname{core} C := \{c \in C: \forall x \in X\ \exists \varepsilon > 0: \forall \lambda \in [-\varepsilon, \varepsilon],\ c + \lambda x \in C\}.$$

In the context of the pair of problems (P), (D), Rockafellar's (CQ) is

$$0 \in \operatorname{core} (\operatorname{dom} g - A \operatorname{dom} f). \tag{R}$$

Robinson's refinement [13, Cor. 1] of a result of Rockafellar [14, Thm. 18] leads to the following theorem.

THEOREM 2.2. *Let* X, Y *be Banach spaces and suppose that* (R) *holds. Then* $\inf (\mathrm{P}) = \max (\mathrm{D})$.

We will show below that the *core* constraint qualification of Rockafellar is the *key* for constructing new constraint qualifications and will, as well, explain most of the classical and more recent constraint qualifications existing in the literature. In particular, we will show that many seemingly unrelated constraint qualifications are in fact related to (R) and show how new duality results can be derived from Theorem 2.2.

3. Comparison of constraint qualifications. In this section we present some constraint qualifications that can be derived from the Rockafellar condition (R). In the first part of this section we discuss the case when A is a continuous linear operator with *finite-dimensional* range, i.e., $A: X \to Y$ with $Y = \mathbb{R}^n$. In the second part we give corresponding results for an operator with infinite-dimensional range.

3.1. A is a linear operator with finite-dimensional range $Y = \mathbb{R}^n$. We first recall the following result, see, e.g., Holmes [9].

[1] Throughout this paper we assume that inf (P) is finite.

PROPOSITION 3.1. *Let X be a real topological vector space and let $C \subset X$ be convex. Then,* int $C \subset$ core C. *Further,* int $C =$ core C *under each of the following conditions:*

(a) int $C \neq \emptyset$.

(b) *X is finite-dimensional.*

From Proposition 3.1 it follows immediately that if int $(\operatorname{dom} g - A \operatorname{dom} f) \neq \emptyset$, then

$$0 \in \operatorname{core}(\operatorname{dom} g - A \operatorname{dom} f) \Leftrightarrow 0 \in \operatorname{int}(\operatorname{dom} g - A \operatorname{dom} f).$$

Recall that for a convex subset $C \subset \mathbb{R}^n$ we have:

$$0 \in \operatorname{int} C \Leftrightarrow \operatorname{cone} C = \mathbb{R}^n. \tag{3.1}$$

where cone $C = \{\lambda x: \lambda \geqq 0, x \in C\}$. Hence,

$$\begin{aligned} 0 \in \operatorname{core}(\operatorname{dom} g - A \operatorname{dom} f) &\Leftrightarrow 0 \in \operatorname{int}(\operatorname{dom} g - A \operatorname{dom} f) \\ &\Leftrightarrow \operatorname{cone}(\operatorname{dom} g - A \operatorname{dom} f) = \mathbb{R}^n. \end{aligned} \tag{3.2}$$

It follows that (R) and (S) are equivalent when $Y = \mathbb{R}^n$.

We recall that for a set C in $\mathbb{R}^n$, $y \in \operatorname{ri} C$ if and only if 0 is an interior point of $C - y$ relative to the affine hull of $(C - y)$. It turns out that

$$0 \in \operatorname{ri} C \quad \text{if and only if cone } C \text{ is a (closed) subspace of } \mathbb{R}^n. \tag{3.3}$$

Thus, in view of (3.2), the constraint qualification

$$0 \in \operatorname{ri}(\operatorname{dom} g - A \operatorname{dom} f) \tag{RR}$$

is weaker than (R). However, the following duality result for a Banach space can be deduced from Theorem 2.2. (The proof is omitted since it is similar to the one given for Theorem 3.5.) For a standard proof see, e.g., [6] or [12].

THEOREM 3.1. *Suppose that X is a locally convex space and $Y = \mathbb{R}^n$. If* (RR) *holds, then* inf (P) = max (D).

The remainder of this subsection is devoted to the comparison of constraint qualifications for *linearly constrained convex programs.* In a recent work, Borwein and Wolkowicz [7] introduced a constraint qualification for the linearly constrained convex program:

$$\inf\{f(x): Ax = b, x \in S\} \tag{L}$$

where S is a convex cone in X, i.e., $S + S \subset S$ and $\lambda S \subset S$ for all $\lambda \geqq 0$ and $b \in \mathbb{R}^n$. The feasible set of (L) is

$$F = \{x \in X: Ax = b, x \in S\}$$

and it is assumed that $F \neq \emptyset$. Note that problem (L) is a special case of problem (P) obtained by replacing f by $f + \delta(\cdot \mid S)$ and g by $\delta(\cdot \mid \{b\})$, where $\delta(\cdot \mid E)$ denotes the indicator function of a given set E. In this setting, the corresponding dual reduces to the concave finite-dimensional problem

$$\sup\{b^T y - (f + \delta(\cdot \mid S))^*(A^* y): y \in \mathbb{R}^n\}. \tag{DL}$$

In what follows, $\overline{\text{cone}}\, E$ stands for the closure of the cone generated by the set E. The following result is proved in [7].

THEOREM 3.2. *Let X be a normed linear space. If*

$$\overline{\operatorname{cone}}(F - S) = X, \tag{BW}$$

then inf (L) = max (DL).

For problem (L), the constraint qualification (R) reads (by (3.2)):

$$\text{cone}\,(b - A(S)) = \mathbb{R}^n.$$

We show below that (BW) is stronger than (R), i.e., (BW)$\Rightarrow$(R). We assume that $A: X \to \mathbb{R}^n$ is *onto*. This assumption is not really restrictive since we can always assume that Y is the range of A, and, after a unitary change, write $Y = \mathbb{R}^n$. Also, we introduce another interesting constraint qualification (to be called (BW)′) related to (BW) and equivalent to (RR). In what follows, the kernel of A is denoted by Ker A.

THEOREM 3.3. *Consider the following constraint qualifications*:

(BW) $$\overline{\text{cone}}\,(F - S) = X,$$

(R) $$0 \in \text{core}\,(b - A(S)),$$

(RR) $$0 \in \text{ri}\,(b - A(S)),$$

(BW)′ $$\text{cone}\,(F - S) + \text{Ker}\,(A) \text{ is a closed subspace of } X.$$

Then (BW)$\Rightarrow$(R)$\Rightarrow$(RR)$\Leftrightarrow$(BW)′.

Proof. (BW)$\Rightarrow$(R): Suppose that (BW) holds. Then,

$$\begin{aligned} \mathbb{R}^n = A(X) &= A(\overline{\text{cone}}\,(F - S)) \\ &\subset \overline{\text{cone}}\,A(F - S) \\ &= \overline{\text{cone}}\,(b - A(S)) \subset \mathbb{R}^n. \end{aligned}$$

So, the closure of the convex set cone $(b - A(S))$ is $\mathbb{R}^n$. A simple separation argument shows that cone $(b - A(S)) = \mathbb{R}^n$. Hence $0 \in$ core $(b - A(S))$ by (3.2).

(R)$\Rightarrow$(RR). The proof follows from (3.2) and (3.3).

(RR)$\Rightarrow$(BW)′. From (3.3) we see that cone $(b - A(S))$ is a (closed) subspace of $\mathbb{R}^n$. Hence

$$\begin{aligned} \text{cone}\,(F - S) + \text{Ker}\,(A) &= A^{-1}[A\ \text{cone}\,(F - S)] \\ &= A^{-1}[\text{cone}\,(b - A(S))] \end{aligned}$$

is a closed subspace of X.

(BW)′$\Rightarrow$(RR). If cone $(F - S) + \text{Ker}\,(A)$ is a subspace of X, then

$$\text{cone}\,(b - A(S)) = A[\text{cone}\,(F - S) + \text{Ker}\,(A)] \text{ is a subspace of } \mathbb{R}^n. \qquad \square$$

In view of the above result, it is clear that Theorem 3.2 is a special case of Theorem 3.1. The following example shows that (R) need not imply (BW) even when X is finite-dimensional.

Example 3.1. Let $X = \mathbb{R}^2$, $S = [-1, 1] \times \{0\}$, $A: \mathbb{R}^2 \to \mathbb{R}$ such that $A(x, y) = x$ and $b = 0$ so that $F = (0, 0)$. Clearly (R) holds since $0 \in$ core $(b - A(S)) =$ int $([-1, 1])$ in $\mathbb{R}$, while $\overline{\text{cone}}\,(F - S) = \mathbb{R} \neq X$.

In a recent work, Borwein and Lewis [6], introduced the notion of *quasi-relative interior*. As we shall see below, this notion is useful in the verification of (RR).

DEFINITION 3.1 [6]. Let X be a topological vector space. For a convex $C \subset X$, the quasi-relative interior of C (qri C) is the set of those $x \in C$ for which $\overline{\text{cone}}\,(C - x)$ is a subspace.

This notion is studied extensively in [6]. For any set E, in finite dimension, $\overline{\text{cone}}\,E$ is a closed subspace if and only if cone E is a subspace; hence the notion of quasi-relative interior coincides with the relative interior. However, what makes the qri useful is the following important property.

PROPOSITION 3.2 [6]. *Suppose* $A: X \to \mathbb{R}^n$ *is a continuous linear map. Then,* $A(\text{qri } C) \subset \text{ri } (AC)$, *and if* $\text{qri } (C) \neq \emptyset$, *then* $A(\text{qri } C) = \text{ri } (AC)$.

Using the above proposition, we see that when $\text{qri } S \neq \emptyset$, the constraint qualification (RR), namely $b \in \text{ri } A(S)$, reads

$$\text{(BL)} \qquad \exists x \in \text{qri } (S) \text{ such that } Ax = b.$$

Thus in view of Theorem 3.1 we have the following theorem.

THEOREM 3.4. *Suppose that* X *is locally convex and* $\text{qri } S \neq \emptyset$. *If* (BL) *holds then* $\inf (\text{L}) = \max (\text{DL})$.

We wish to emphasize the importance of this result. The importance lies in the fact that in many applications (for example, when S is the nonnegative cone L_p^+, $1 \leqq p < \infty$), $\text{qri } S \neq \emptyset$ while $\text{ri } S = \emptyset$.

3.2. A is a linear operator with infinite-dimensional range. We have shown in the previous subsection that the core condition provides a systematic mechanism for constructing old and new constraint qualifications. The natural question is now to see whether similar results can be derived for the general case. Unless otherwise specified, in the sequel we assume that X and Y are Banach spaces. We recall that for a convex subset C of an infinite-dimensional vector space X:

$$0 \in \text{core } C \Leftrightarrow \text{cone } C = X \quad \text{and} \quad 0 \in \text{int } C \Rightarrow \text{cone } C = X.$$

When $A: X \to Y$ with $Y = \mathbb{R}^n$, we were able to relax (R) by (RR) using the notion of relative interior instead of that of interior and core. Following the same methodology, by introducing the concept of *intrinsic core* we may establish an appropriate (CQ) when Y is a Banach space.

DEFINITION 3.2 [9]. The core of C relative to aff C, the affine hull of C, is called the intrinsic core of C and is written icr C.

When $C \subset X$ is convex and X is finite-dimensional we have

$$\text{icr } C = \text{ri } C.$$

Recall that aff C is $x + \text{span } (C - x)$ for any fixed $x \in C$, where span $(C - x)$ is the smallest subspace of X that contains $(C - C)$.

PROPOSITION 3.3. *Let* C *be a convex subset of* X. *Then,*

$$x \in \text{icr } C \Leftrightarrow \text{cone } (C - x) = \text{aff } (C - x) = \text{aff } (C - C).$$

Proof. By Definition 3.2, we have $x \in \text{icr } C \Leftrightarrow \text{cone } (C - x) = \text{aff } (C - x)$. But, when $x \in C$, $\text{aff } (C - x) = \text{aff } ((C - x) - (C - x)) = \text{aff } (C - C)$. □

In the finite-dimensional setting, $x \in \text{ri } C$ if and only if $\text{cone } (C - x) = \text{aff } (C - C)$ and further, aff $(C - C)$ is closed. Thus, a natural (CQ) in a general setting should now be:

$$x \in \text{icr } C \quad \text{and} \quad \text{aff } (C - C) \quad \text{is a closed subspace.}$$

For the problem (P) this general constraint qualification reads:

$$\text{(GCQ)} \qquad 0 \in \text{icr } (\text{dom } g - A \text{ dom } f) \quad \text{and}$$

$$\text{aff } (\text{dom } g - A \text{ dom } f) \quad \text{is a closed subspace.}$$

From Theorem 2.2, we know that a strong duality result is guaranteed if

$$0 \in \text{core } (\text{dom } g - A \text{ dom } f).$$

Since $(\operatorname{dom} g - A \operatorname{dom} f)$ is a convex subset of Y, condition (R) is equivalent to:

$$\operatorname{cone}(\operatorname{dom} g - A \operatorname{dom} f) = Y.$$

In the context of problem (P), if $0 \in \operatorname{core}(\operatorname{dom} g - A \operatorname{dom} f)$, then

$$0 \in \operatorname{icr}(\operatorname{dom} g - A \operatorname{dom} f) \quad \text{and} \quad \operatorname{aff}(\operatorname{dom} g - A \operatorname{dom} f) = Y \text{ (a closed subspace).}$$

Thus, the constraint qualification (R) is stronger than the constraint qualification (GCQ). However, as we see below, the strong duality result under (GCQ) can be deduced from Theorem 2.2. In this sense, (GCQ) and (R) are equivalent.

THEOREM 3.5. *Let X, Y be Banach spaces and let $A: X \to Y$ be a continuous map. Let $f: X \to (-\infty, +\infty]$ and let $g: Y \to (-\infty, +\infty]$ be proper, lower semicontinuous convex functions. Suppose that* (GCQ) *holds, i.e.,*

$$0 \in \operatorname{icr}(\operatorname{dom} g - A \operatorname{dom} f) \quad \textit{and} \quad \operatorname{aff}(\operatorname{dom} g - A \operatorname{dom} f) \quad \textit{is a closed subspace.}$$

Then, $\inf(\mathrm{P}) = \max(\mathrm{D})$.

Proof. Let $x_0 \in \operatorname{dom} f$ such that $Ax_0 \in \operatorname{dom} g$. Define

$$F(x) := f(x + x_0) \quad \text{and} \quad G(y) := g(y + Ax_0).$$

Then, we have $\operatorname{dom} F = \operatorname{dom} f - x_0$, $\operatorname{dom} G = \operatorname{dom} g - Ax_0$,

$$F^*(x^*) = f^*(x^*) - \langle x^*, x_0 \rangle, \qquad G^*(y^*) = g^*(y^*) - \langle y^*, Ax_0 \rangle,$$

$$\inf_{x \in X} \{f(x) + g(Ax)\} = \inf_{x \in X} \{F(x) + G(Ax)\},$$

and

$$\sup_{y^* \in Y^*} \{-g^*(y^*) - f(-{}^*y^*)\} = \sup_{y^* \in Y^*} \{-G^*(y^*) - F^*(-A^*y^*)\}.$$

Further, if in the last equation, sup is attained in the right-hand side, then sup is attained in the left-hand side. Also,

$$\operatorname{dom} G - A(\operatorname{dom} F) = \operatorname{dom} g - A(\operatorname{dom} f).$$

We see that $F(0) = f(x_0)$ and $G(0) = g(Ax_0)$ are real numbers. Thus, without loss of generality, we can assume that

$$0 \in \operatorname{dom} f \quad \text{and} \quad 0 \in \operatorname{dom} g. \tag{3.4}$$

Since (GCQ) holds, by Proposition 3.3,

$$M := \operatorname{cone}(\operatorname{dom} g - A \operatorname{dom} f) = \operatorname{aff}(\operatorname{dom} g - A \operatorname{dom} f).$$

It is given that M is a closed subspace of Y. Then $\hat{X} = A^{-1}(M)$ is a Banach space, $\operatorname{dom} f \subset \hat{X}$, $\operatorname{dom} g \subset M$ from (3.4). We replace X by $\hat{X}$, Y by M in problem (P) and regard A as a mapping from $\hat{X} \to M$. For the corresponding pair of transformed problems (P′), (D′)

$$(\mathrm{P}') \qquad \inf_{x \in \hat{X}} \{f(x) + g(Ax)\}$$

$$(\mathrm{D}') \qquad \sup_{y^* \in M^*} \{-g^*(y^*) - f^*(-A^*y^*)\}$$

the condition (R) holds, namely,

$$0 \in \operatorname{core}(\operatorname{dom} g - A \operatorname{dom} f).$$

By Theorem 2.2, we see that $\inf(P') = \max(D')$. To complete the proof it remains to show that $\inf(P') = \inf(P)$ and $\max(D') = \max(D)$. Clearly, since $\operatorname{dom} f \subset \hat{X}$, it follows that $\inf(P') = \inf(P)$.

Let $v^* \in M^*$ be such that

$$\max(D') = -g^*(v^*) - f^*(-A^*v^*).$$

For any $y^* \in Y^*$, let Γy^* denote the restriction of y^* to M. It is easily seen that

$$g^*(y^*) = g^*(\Gamma y^*) \quad (\text{since } \operatorname{dom} g \subset M)$$

and

$$f^*(A^*(\Gamma y^*)) = f^*(A^*y^*) \quad (\text{since } A\hat{X} \subset M).$$

Therefore,

$$\sup_{y^* \in Y^*} \{-g^*(y^*) - f^*(-A^*y^*)\} = \sup_{y^* \in Y^*} \{-g^*(\Gamma y^*) - f^*(-A^*(\Gamma y^*))\}$$

$$= \sup_{w^* \in \Gamma(Y^*)} \{-g^*(w^*) - f^*(-A^*w^*)\}.$$

By the Hahn-Banach extension theorem, $\Gamma(Y^*) = M^*$ and hence

$$\sup(D) = \sup_{w^* \in M^*} \{-g^*(w^*) - f^*(-A^*w^*)\} = \sup(D') = -g^*(v^*) - f^*(-A^*v^*).$$

But it is clear that $\sup(D)$ is attained by any continuous linear extension of v^* to Y. Hence the above equality gives $\max(D) = \max(D')$. □

A proof of the above result for the special case, $X = Y$ and $A =$ Identity, appears in Attouch and Brezis [2]. The proof there is based on the Banach-Dieudonne-Krein-Smulian theorem [8, Thm. V.5.7]. Based on this special case, using the notion of strong quasi-relative interior (see Definition 3.3 below), Borwein et al. [5] prove Theorem 3.5. Using a completely different approach, Zalinescu [15, Cor. 4] shows that the above theorem of Attouch and Brezis is valid when X and Y are Fréchet spaces. It is important to note that, by modifying the argument of [5, Thm. 3.1], *Theorem* 3.5 *will remain valid when X and Y are Fréchet spaces.* At this juncture, we wish to mention an earlier work with applications to perturbational duality by Borwein [4]. We thank one of the referee's for bringing this reference to our attention.

In [5], the notion of strong quasi-relative interior is introduced as a natural extension of the quasi-relative interior.

DEFINITION 3.3. For a convex subset $C \subset X$, the strong quasi-relative interior of C is the set of those $x \in C$ for which $\operatorname{cone}(C - x)$ is a closed subspace.

When X is finite-dimensional we have

$$\operatorname{sqri} C = \operatorname{ri} C = \operatorname{qri} C = \operatorname{icr} C.$$

In the context of problem (P), the following (CQ) is proposed in [5]:

$$0 \in \operatorname{sqri}(\operatorname{dom} g - A \operatorname{dom} f).$$

As the following proposition shows, the above constraint qualification given in terms of the strong quasi-relative interior is equivalent to the constraint qualification (GCQ).

PROPOSITION 3.4.

$$\left\{ \begin{array}{c} x \in \operatorname{icr}(C) \\ \operatorname{aff}(C - x) \text{ is a closed subspace} \end{array} \right\} \Leftrightarrow x \in \operatorname{sqri} C.$$

Proof. If $x \in \text{icr } C$ and $\text{aff}(C-x)$ is a closed subspace, then by Proposition 3.3 $\text{cone}(C-x) = \text{aff}(C-x)$ is a closed subspace, and hence $x \in \text{sqri } C$ from Definition 3.3. On the other hand, if $x \in \text{sqri } C$, then $\text{cone}(C-x)$ is a closed subspace. But, in this situation, $\text{cone}(C-x) \subset \text{aff}(C-x)$ and $\text{aff}(C-x) \subset \text{cone}(C-x)$. Hence $\text{aff}(C-x) = \text{cone}(C-x)$ and thus $x \in \text{icr } C$ by Proposition 3.3. □

The name strong quasi-relative interior (sqri) has been introduced as a natural generalization of the quasi-relative interior. However, our results below demonstrate that in fact the strong quasi-relative interior is "closer" to the relative interior. Recall that for a set C in a topological space X, $y \in \text{ri } C$ if and only if 0 is an interior point of $C-y$ relative to the closure of the affine hull of $C-y$, see [11].

We prove the following results in the general setting of Baire spaces. Recall that X is a Baire space if it is locally convex and the intersection of every countable collection of dense open subsets of X is dense in X. Every closed subspace of such a space is Baire and such a space is barrelled, i.e., each absorbing, convex, circled, and closed subset of X is a neighborhood of the origin. Examples of Baire spaces are Fréchet spaces and Banach spaces (see [10]).

THEOREM 3.6. *Let X be a Baire space and C be a closed convex set in X. Then*

$$\text{sqri } C \subset \text{ri } C.$$

Proof. If $\text{sqri } C = \emptyset$, then there is nothing to prove. Let $\hat{x} \in \text{sqri } C$ so that $Y := \text{cone}(C-\hat{x})$ is a closed subspace of X. Let $K := C-\hat{x}$. Note that $0 \in K$ and K is absorbing in Y. Let $B = \bigcap_{|\lambda| \geq 1} \lambda K$ be the balanced core of K (see [10, p. 80]). We note that

(i) $0 \in B \subset K$,

(ii) B is balanced, closed convex in Y,

(iii) B is absorbing in Y.

Statement (i) follows immediately from the definition of B and (ii) follows since each λK is closed and convex. To see (iii), let $y \in Y$. Since K is absorbing we can find $\mu > 0$ such that $\pm \mu y \in K$. Then from the convexity of K it follows that for every $|\lambda| \geq 1$, $\mu y/\lambda \in K$, and so $\mu y \in B$.

Since X is Baire, Y is also Baire and hence barrelled. B, being an absorbing, balanced, closed, and convex set in Y, is a neighborhood of 0 in Y, and hence

$$0 \in \text{int}_Y B \subset \text{int}_Y K = \text{int}_Y (C-\hat{x})$$

where int_Y denotes the interior relative to Y. Therefore $\hat{x} \in \text{ri } C$ and the proof is complete. □

We remark that the above result may not hold for barrelled spaces since a closed subspace of a barrelled space need not be barrelled.

COROLLARY 3.1. *Let E be a convex set in X where X is a Baire space. Then*

$$\text{sqri } E \subset \text{ri } \bar{E}.$$

Proof.

$$\begin{aligned} \hat{x} \in \text{sqri } E &\Rightarrow \text{cone}(E-\hat{x}) =: Y \text{ is a closed subspace} \\ &\Rightarrow \text{cone}(\overline{E-\hat{x}}) = Y \text{ (closure with respect to } X) \\ &\Rightarrow \text{cone}(\bar{E}-\hat{x}) = Y \\ &\Rightarrow \hat{x} \in \text{sqri } \bar{E} \subset \text{ri } \bar{E}. \end{aligned}$$

Since $\bar{E}$ is closed convex in X, the last inclusion follows from the previous theorem. □

We have seen, as a consequence of Proposition 3.4, that the constraint qualification expressed in terms of the strong quasi-relative interior is equivalent to (GCQ). The above corollary suggests looking at the following weaker condition to guarantee the strong duality result:

$$0 \in \operatorname{ri} \overline{(\operatorname{dom} g - A \operatorname{dom} f)}. \tag{3.5}$$

The following examples demonstrate that

(i) equality may not hold in Corollary 3.1, and

(ii) the strong duality result may not hold under (3.5).

Example 3.2. Let X be an infinite-dimensional Banach space. Let $\phi : X \to \mathbb{R}$ be a noncontinuous linear functional so that $S := \operatorname{Ker} \phi$ is a dense subspace of X. For any $e \in X \setminus S$, we see that

$$X = S + \mathbb{R}e \quad \text{and} \quad S \cap \mathbb{R}e = \{0\}.$$

Let $E := S + [0, 1]e$ where $[0, 1] = \{\lambda : 0 \leqq \lambda \leqq 1\}$. Clearly, $\bar{E} \supset \bar{S} = X$ and hence $\bar{E} = X$, so that $\operatorname{ri} \bar{E} = \operatorname{ri} X = \operatorname{int} X = X$ contains 0.

To get a contradiction, suppose that $0 \in \operatorname{sqri} E$. Then $0 \in \operatorname{icr} E$, i.e., $0 \in \operatorname{core} E$ relative to aff E. Now $-e \in \operatorname{aff} E$ and hence there exists $\lambda > 0$ such that

$$-\lambda e \in E = S + [0, 1]e.$$

Thus, $-\lambda e = s + \mu e$ for some $s \in S$ and $\mu \in [0, 1]$. This implies that $e \in S$, a contradiction. Thus $0 \in \operatorname{ri} \bar{E}$ while $0 \notin \operatorname{sqri} E$. We note that aff $E = X$ is closed while $0 \notin \operatorname{icr} E$.

Example 3.3. As in [11, p. 77] we consider the following setting:

$$X = l_2 = \left\{ x = (x_1, \cdots, x_n, \cdots) : x_n \in \mathbb{R}, \sum_1^\infty x_n^2 < \infty \right\},$$

$$C = \{x \in l_2 : x_{2n-1} + x_{2n} = 0, \forall n = 1, 2, \cdots,\},$$

$$S = \{x \in l_2 : x_{2n} + x_{2n+1} = 0, \forall n = 1, 2, \cdots,\}.$$

Clearly, C and S are closed subspaces of X and $C \cap S = \{0\}$. Define f and g on X by $f(x) = \delta(x \mid C)$ and $g(x) = x_1$ if $x \in S$ and ∞ otherwise. It is easily seen that f and g are convex and lower semicontinuous on X with $\operatorname{dom} f = C$ and $\operatorname{dom} g = S$. We now compute the conjugates of f and g. Since C is a subspace it is easy to see that

$$f^*(x^*) = \delta(x^* \mid C^\perp)$$

where $C^\perp$ is the orthogonal complement of C. Also we have,

$$\begin{aligned} g^*(x^*) &= \sup_{x \in S} \{\langle x, x^* \rangle - x_1\} \\ &= \sup_{x \in S} \langle x^* - e_1, x \rangle \quad (\text{where } e_1 = (1, 0, \cdots)) \\ &= \begin{cases} 0 & \text{if } x^* - e_1 \in S^\perp \\ \infty & \text{if } x^* - e_1 \notin S^\perp \end{cases} \\ &= \delta(x^* \mid e_1 + S^\perp). \end{aligned}$$

We claim that the following are true:

(i) $0 \in \operatorname{ri} \overline{(\operatorname{dom} g - \operatorname{dom} f)}$,

(ii) $0 \notin \operatorname{sqri} (\operatorname{dom} g - \operatorname{dom} f)$,

(iii) $\inf_{x \in X} \{f(x) + g(x)\} = 0$,

(iv) $\sup_{x^* \in X^*} \{-g^*(x^*) - f^*(-x^*)\} = -\infty$.

It follows from these that the strong duality result fails to hold under the weaker constraint qualification

$$0 \in \operatorname{ri} \overline{(\operatorname{dom} g - \operatorname{dom} f)}.$$

To see (i), we show that $(\operatorname{dom} g - \operatorname{dom} f) = S - C$ is dense in X. To this end, let $x = (x_n)$ be orthogonal to $S - C$. Since $e_{2n-1} - e_{2n} \in C$ and $e_{2n} - e_{2n+1} \in S$ for all $n = 1, 2, \cdots$, we see that $x_{2n-1} - x_{2n} = 0$ and $x_{2n} - x_{2n+1} = 0$ for all n. Since $x \in l_2$ we must have $x = 0$ so that $S - C$ is dense in X.

Statement (ii) follows immediately from the observation that

$$\operatorname{aff}(\operatorname{dom} g - \operatorname{dom} f) = S - C \quad \text{is not closed.}$$

Note however that $0 \in \operatorname{icr}(\operatorname{dom} g - \operatorname{dom} f)$.

Now $\inf_{x \in X} \{f(x) + g(x)\} = \inf_{x \in \operatorname{dom} g \cap \operatorname{dom} f} = f(0) + g(0) = 0$ gives (iii). We now show that

$$\operatorname{dom} g^* \cap \operatorname{dom} f^* = \varnothing \tag{3.6}$$

so that

$$\sup_{x^* \in X^*} \{-g(x^*) - f^*(-x^*)\} = \sup_{x^* \in \operatorname{dom} g^* \cap \operatorname{dom} f^*} \{-g(x^*) - f^*(-x^*)\} = -\infty$$

giving (iv). To see (3.6), suppose that

$$(x_n) = x \in \operatorname{dom} g^* \cap \operatorname{dom} f^* = (e_1 + S^\perp) \cap C^\perp.$$

Then, as in the proof of (i), we get $x_{2n-1} - x_{2n} = 0$ and $x_{2n} - x_{2n+1} = 0$ for all $n = 1, 2, \cdots$. Hence, $x = 0$. But then $0 \in e_1 + S^\perp$ implies $-e_1 \in S^\perp$, which is false since $e_1 \in S$.

Our last result resembles Proposition 3.2 and partially addresses the question of verifying (GCQ).

PROPOSITION 3.5. *Let X be a locally convex topological vector space and let Y be a Baire space. Let $A: X \to Y$ be a continuous linear operator and C be a convex set in X. Then*

$$\operatorname{sqri} A(C) \subset \overline{A(\operatorname{qri} C)}$$

whenever $\operatorname{qri} C \neq \varnothing$.

Proof. From Corollary 3.1 we have

$$\operatorname{sqri} A(C) \subset \operatorname{ri} \overline{A(C)}.$$

Let $x_1 \in \operatorname{qri} C$ and $y \in \operatorname{ri} \overline{A(C)}$. Since y and Ax_1 belong to $\overline{A(C)}$, we have for some $\varepsilon > 0$, $\varepsilon(y - Ax_1) \in \overline{A(C)} - y$, i.e.,

$$y - \varepsilon(Ax_1 - y) \in \overline{A(C)}.$$

Let V be any convex, balanced neighborhood of 0. Then there exists $u \in V$ such that

$$y - \varepsilon(Ax_1 - y) + u = Ax_2 \quad \text{for some } x_2 \in C$$

and thus

$$y + \frac{u}{1+\varepsilon} = \frac{Ax_2 + \varepsilon Ax_1}{1+\varepsilon}$$

from which it follows that

$$y + \frac{u}{1+\varepsilon} \in A(\operatorname{qri} C)$$

since $(x_2 + \varepsilon x_1)/(1+\varepsilon) \in \operatorname{qri} C$ by [6, Lemma 2.9]. Now

$$u/(1+\varepsilon) \in V \quad \text{and hence} \quad (y + V) \cap A \operatorname{qri} C \neq \varnothing$$

implying that $y \in \overline{A(\operatorname{qri} C)}$. □

REFERENCES

[1] J. P. AUBIN, *Mathematical Methods of Game and Economic Theory*, North-Holland, Amsterdam, 1979.

[2] H. ATTOUCH AND H. BREZIS, *Duality for the sum of convex functions in general Banach Spaces*, in Aspects of Mathematics and its Applications, J. A. Barroso, ed., Elsevier Science, Amsterdam, 1986.

[3] A. BEN-TAL, J. M. BORWEIN, AND M. TEBOULLE, *A dual approach to multi-dimensional L_p spectral estimation problems*, SIAM J. Control Optim., 26 (1988), pp. 985-996.

[4] J. M. BORWEIN, *Convex relations in analysis and optimization*, in Generalized Concavity in Optimization and Economics, S. Shaible and W. T. Ziemba, eds., Academic Press, New York, 1981, pp. 335-377.

[5] J. M. BORWEIN, V. JEYAKUMAR, A. LEWIS, AND H. WOLKOWICZ, *Constrained approximation via convex programming*, preprint, 1988.

[6] J. M. BORWEIN AND A. S. LEWIS, *Partially finite convex programming*, Tech. Report, Dalhousie University, 1988.

[7] J. M. BORWEIN AND H. WOLKOWICZ, *A simple constraint qualificaty in infinite dimensional programming*, Math. Programming, 35 (1986), pp. 83-96.

[8] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators Part I*, Interscience, New York, 1964.

[9] R. B. HOLMES, *Geometric Functional Analysis and Its Applications*, Springer-Verlag, New York, 1975.

[10] J. HORVATH, *Topological Vector Spaces and Distributions, Volume I*, Addison-Wesley, Reading, MA, 1966.

[11] J. PONSTEIN, *Approaches to the Theory of Optimization*, Cambridge University Press, Cambridge, 1980.

[12] P. J. LAURENT, *Approximation et Optimisation*, Herman, Paris, 1972.

[13] S. M. ROBINSON, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130-143.

[14] R. T. ROCKAFELLAR, *Conjugate Duality and Optimization*, CBMS-NSF Regional Conference In Applied Mathematics 16, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.

[15] C. ZALINESCU, *Solvability results for sublinear functions and operators*, Z. Oper. Res., 31 (1987), pp. A79-A101.

[16] J. ZOWE AND KURCYUSZ, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49-62.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 936-949, July 1990

009

SOLUTION AND CONTROL OF A BILINEAR STOCHASTIC DELAY EQUATION*

FRANCO FLANDOLI†

Abstract. A bilinear stochastic delay equation with delay in the coefficients of the noise is considered. When the equation is rewritten in abstract form in Hilbert spaces, the coefficients of the noise are unbounded operators. A direct solution of the abstract equation is presented, under a general assumption which allows unification of this class of delay equations with a class of parabolic equations. Finally, the linear quadratic regulator problem governed by this equation is studied by a direct solution of the associated Riccati equation.

Key words. bilinear stochastic delay equation, optimal control, Riccati equation

AMS(MOS) subject classifications. 49B60, 60H20, 93E20

1. Introduction.

Consider the following linear stochastic functional differential equation:

$$x(t) = y^0 + \int_0^t F^0(x_s)\, ds + \sum_{j=1}^m \int_0^t F^j(x_s)\, dw^j(s) + \int_0^t B^0 u(s)\, ds, \tag{1.1}$$

$$t \in [0, T], \quad x_0 = y^1.$$

Here $x(t)$ is a R^d-valued stochastic process, r is a fixed positive real number, $x_t: [-r, 0] \to R^d$ denotes the function $x_t(s) = x(t+s)$, $w = (w^1, \cdots, w^m)$ is a standard m-dimensional Brownian motion on a complete probability space (Ω, F, P), $F^0, F^1, \cdots, F^m$ are bounded linear operators from $C([-r, 0]; R^d)$ into R^d, $u(t)$ is a R^k-valued stochastic process (the control function), $B^0 \in R^{k \times d}$, and finally $(y^0, y^1) \in Y$, where Y is the Hilbert space $R^d \times L^2(-r, 0; R^d)$.

Our final purpose is to study an optimal control problem for (1.1). To this end it is convenient to rewrite the concrete delay equation as an abstract stochastic equation in the Hilbert space Y of the form

$$y(t) = S(t)y + \sum_{j=1}^m \int_0^t S(t-s) D^j y(s)\, dw^j(s) + \int_0^t S(t-s) B u(s)\, ds \tag{1.2}$$

(the strongly continuous semigroup $S(t)$ in Y, and the operators D^j and B, will be defined in § 2.1). An important feature of (1.2) is the unboundedness of the operators D^j (corresponding to F^j).

Let us first consider the problem of the well-posedness of (1.2) and related closed loop equations. Although a solution of (1.2) is $y(t) = (x(t), x_t)$, where $x(t)$ is the solution of the concrete equation (1.1) (this approach is briefly discussed in § 2), it is of interest to directly consider (1.2) from an abstract point of view, and to see if it is possible to identify some general assumptions, possibly common to other problems, which are sufficient to study this equation directly. Note that classical results (see for instance [I.1]) cannot be applied, because of the unboundedness of the operators D^j.

A partially analogous problem was studied by Da Prato [D.1] (see also [D.2] and [D-I.1]) in the context of stochastic partial differential equations of parabolic type. In [D.1], $S(t)$ is generated by a second-order elliptic operator (in the typical situation),

* Received by the editors January 18, 1989; accepted for publication (in revised form) October 12, 1989.
† Dipartimento di Matematica, Universita' di Torino, Via Principe Amedeo 8, 10123 Torino, Italy.

and the operators D^j correspond to first-order differential operators. Hence D^j are unbounded operators, as in the present paper. Besides this formal similarity, it is possible to identify a basic property which is common to these two problems. In the case studied by [D.1], under a natural coercivity assumption, there exists a constant $c \in (0, 1)$ such that

$$\sum_{j=1}^{m} \int_0^t |D^j S(t)y|^2 \, dt \leqq c|y|^2$$

for every $y \in Y$, where $|\cdot|$ denotes the norm in Y. Although the solution method of [D.1] is not based on this property but on a similar result for stochastic convolution integrals, we can show that this inequality is in fact sufficient to solve (1.2), in the case of [D.1]. The remarkable fact is that a similar inequality can be proved for problem (1.1). More precisely, if we use the natural norm of Y we cannot have $c < 1$ in general (unless artificial assumptions on $F^0, \cdots, F^m$ are imposed), but for every $c' \in (0, 1)$ we can find an equivalent norm in Y, denoted by $|\cdot|_\lambda$ (see § 2.2) such that

$$\sum_{j=1}^{m} \int_0^\tau |D^j S(t)y|_\lambda^2 \, dt \leqq (c' + \tau c'')|y|_\lambda^2$$

for every $\tau \in [0, T]$ and $y \in Y$, and for some constant $c'' > 0$. This inequality allows us to unify the two problems and solve (1.2) directly (see § 3).

Finally, in § 4 we study an optimal control problem over finite time horizon, with quadratic cost functional, for (1.1) (or (1.2)). The solution of this problem is based on the following Riccati equation, written in mild form with inner products:

$$\begin{aligned} \langle P(t)x, y\rangle = {} & \langle P_T S(T-t)x, S(T-t)y\rangle + \sum_{j=1}^{m} \int_t^T \langle P(s)D^j S(s-t)x, D^j S(s-t)y\rangle \, ds \\ & + \int_t^T \langle C(s)S(s-t)x, C(s)S(s-t)y\rangle_Z \, ds \\ & - \int_t^T \langle N(s)^{-1}B^* P(s)S(s-t)x, B^* P(s)S(s-t)y\rangle_k \, ds \end{aligned} \tag{1.3}$$

for every $x, y \in Y$. Here and in the sequel $\langle \cdot, \cdot \rangle$ denotes the inner product in Y, while the other inner products are explicitly indicated; moreover, the operators P_T, $C(s)$, $N(s)$, and the space Z are introduced in § 4.

In the direct solution of the Riccati equation (1.3) we also meet the problem of the unboundedness of D^j. Moreover, the Riccati equation, unlike (1.2), explicitly depends on the norm chosen a priori in Y, via the cost functional. However, by a simple linear transformation it is possible to rewrite (1.3) (as well as the cost functional) with respect to the inner product $\langle \cdot, \cdot \rangle_\lambda$. Thus we can use the basic inequality mentioned above to study (1.3) directly by a standard contraction principle. For the parabolic problem of [D.2] a direct solution of (1.3) is presented in [D-I.1]. Finally, in § 4.4 we solve the synthesis of the associated optimal control problem.

1.1. Notation. We denote the norm and inner product in Y by $|\cdot|$ and $\langle \cdot, \cdot \rangle$, respectively, and the norm and inner product in R^n by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle_n$, the corresponding value of n (in the case of the norm) being clear by the context. Other norms and inner products will be explicitly indicated.

Given a separable Hilbert space X, we denote by $M_w^2(0, T; X)$ the space of all processes in $L^2(0, T; L^2(\Omega; X))$ adapted to $w(t)$, and by $C_w(0, T; X)$ the subspace of processes in $C([0, T]; L^2(\Omega; X))$. Moreover, we denote by $\Sigma(X)$ the space of all bounded self-adjoint operators in X, and by $\Sigma^+(X)$ the subspace of those that are nonnegative definite.

Given two Banach spaces X and Y, we denote by $L(X, Y)$ the space of all bounded linear operators from X to Y, and we set $L(X) = L(X, X)$. If A is a linear operator from X to Y, not necessarily bounded, we denote by $\mathcal{D}(A)$ its domain. Moreover we denote by $B(0, T; L(X, Y))$ the Banach space of all $P(\cdot):[0, T] \to L(X, Y)$ which are strongly measurable and uniformly bounded over $[0, T]$, i.e., such that $\sup_{0 \leqq t \leqq T} |P(t)|_{L(X,Y)} < \infty$.

2. Properties of the stochastic delay equation (1.1).

2.1. Abstract formulation of the stochastic delay equation. This introductory section is devoted to a brief discussion of the concrete equation (1.1) and to its reformulation in the abstract form (1.2). The results are standard, or can be proved by standard arguments, so we will only sketch the proofs.

We recall that the operators F^j in (1.1) are assumed to be bounded and linear from $C([-r, 0]; R^d)$ into R^d. Although F^j are only defined on continuous functions, the quantities $F^j(x_t)$ still make sense as functions of t with values in R^d, for every $x(\cdot)$ in $L^2(-r, T; R^d)$. Indeed we have the following lemma.

LEMMA 2.1. *If $x(\cdot) \in L^2(-r, T; R^d)$ and F is a bounded linear operator from $C([-r, 0]; R^d)$ into R^d, then the function $t \to F(x_t)$ belongs to $L^2(0, T; R^d)$, and there exists a constant $c > 0$ (independent of $x(\cdot)$) such that*

$$\int_0^T \|F(x_t)\|^2 \, dt \leqq c \int_{-r}^T \|x(t)\|^2 \, dt. \tag{2.1}$$

The standard proof is based on the representation of $F(f)$ in the form $F(f) = \int_{-r}^0 d\eta(s) f(s)$, where $\eta(\cdot)$ is a $d \times d$ matrix of functions of bounded variation (cf. [D-S.1, p. 240]): inequality (2.1) is first proved for $x \in C([-r, T]; R^d)$, by Fubini's theorem; thus (2.1) along with the density of $C([-r, T]; R^d)$ into $L^2(-r, T; R^d)$ provides the definition of $t \to F(x_t)$ when $x \in L^2(-r, T; R^d)$, and (2.1) holds true also for this larger class.

Using the previous lemma, along with a standard application of Doob's inequality and contraction principle (or successive approximations), we easily obtain the following result (see for instance [M.1]).

THEOREM 2.2. *Let $u \in M_w^2(0, T; R^k)$ and $(y^0, y^1) \in Y$. Then there exists a unique continuous stochastic process $x \in L^2(\Omega; C([0, T]; R^d))$, adapted to $w(t)$, satisfying* (1.1). *Moreover, there exists a positive constant c independent of u, y^0, and y^1, such that*

$$E \sup_{0 \leqq t \leqq T} \|x(t)\|^2 \leqq c|(y^0, y^1)|^2 + cE \int_0^T \|u(s)\|^2 \, ds. \tag{2.2}$$

We now reformulate (1.1) in the abstract form (1.2). Define the Y-valued process $y(t)$ as

$$y(t) = (x(t), x_t)$$

where $x(t)$ is the solution of (1.1). Moreover, define the linear operators $A: \mathcal{D}(A) \subset Y \to Y$, $D^j: \mathcal{D}(D^j) \subset Y \to Y$, and $B: R^k \to Y$ as follows:

$$\mathcal{D}(A) = \{(y^0, y^1) \in Y: y^1 \in H^1(-r, 0; R^d),\ y^1(0) = y^0\},$$

$$A(y^0, y^1) = (F^0(y^1), dy^1/dt);$$

$$\mathcal{D}(D^j) = R^d \times C([-r, 0]; R^d), \quad D^j(y^0, y^1) = (F^j(y^1), 0), \qquad j = 1, \cdots, m;$$

$$Bu = (B^0 u, 0) \quad \text{for every } u \in R^k.$$

It is well known that A generates a strongly continuous semigroup in Y (cf. [D-M.1]), denoted here by $S(t)$.

THEOREM 2.3. *$y(\cdot)$ is a continuous adapted Y-valued stochastic process belonging to $L^2(\Omega; C([0, T]; Y))$, and satisfying the mild equation* (1.2), *with $y = (y^0, y^1)$.*

Proof. The proof that $y(\cdot)$ is a continuous adapted Y-valued process is standard; we only explicitly note that the second component of y is a continuous $L^2(-r, 0; R^d)$-valued process, because almost surely $x \in L^2(-r, T; R^d)$, whence

$$\lim_{t \to t_0} \int_{-r}^{0} \|x(t+s) - x(t_0+s)\|^2 \, ds = 0 \quad \text{a.s.},$$

for every $t_0 \in [0, T]$. Moreover, $x_\cdot \in L^2(\Omega; C([0, T]; L^2(-r, 0; R^d)))$ because

$$\int_{-r}^{0} \|x_t(s)\|^2 \, ds \leqq \int_{-r}^{T} \|x(s)\|^2 \, ds.$$

($x_\cdot$ is a $C([0, T]; L^2(-r, 0; R^d))$-valued random variable since the Borel σ-field of this space is generated by the evaluations.)

Let us now show that $y(\cdot)$ satisfies (1.2). Recall that $S(t)y = (x^*(t), x_t^*)$, where $x^*(t)$ is the unique solution of the homogeneous equation

$$x^*(t) = y^0 + \int_0^t F^0(x_s^*) \, ds, \qquad t \in [0, T], \qquad x_0^* = y^1. \tag{2.3}$$

If $U(t)$ denotes the fundamental solution associated to this problem, i.e., the $R^{d \times d}$-valued function satisfying

$$U(t) = I + \int_0^t F^0(U_s) \, ds, \qquad U_0 = 0$$

(I is the identity in R^d), then $S(t)(y^0, 0) = (U(t)y^0, U_t y^0)$. With this notation let us define the process $\bar{x}(t)$ as

$$\bar{x}(t) = x^*(t) + \sum_{j=1}^{m} \int_0^t U(t-s) F^j(x_s) \, dw^j(s) + \int_0^t U(t-s) B^0 u(s) \, ds, \tag{2.4}$$

$$t \in [0, T], \quad \bar{x}_0 = y^1.$$

Clearly $\bar{x}(\cdot) \in C_w(0, T; R^d)$ (while the sample continuity of $\bar{x}(t)$ is less obvious a priori). From (2.4) it follows

$$\bar{x}_t = x^* + \sum_{j=1}^{m} \int_0^t U_{t-s} F^j(x_s) \, dw^j(s) + \int_0^t U_{t-s} B^0 u(s) \, ds. \tag{2.5}$$

Indeed, for $v \in [-r, 0]$, the integral $\int_0^t U(t+v-s) F^j(x_s) \, dw^j(s)$ is equal to $\int_0^{t+v} U(t+v-s) F^j(x_s) \, dw^j(s)$ if $t+v \geqq 0$, and to 0 if $t+v < 0$, because $U(t+v-s) = 0$ for $s \in [t+v, t]$; a similar identity holds for the last integral in (2.4), so that (2.5) is proved.

If we prove that $y(t)=(\bar{x}(t), \bar{x}_t)$, then we obtain that $y(\cdot)$ satisfies (1.2), by virtue of (2.4) and (2.5). Thus we have to prove that almost surely $x=\bar{x}$ (in the sense of functions of $L(-r, T; R^d)$). We have

$$\begin{aligned}\int_0^t F^0(\bar{x}_s)\,ds &= \int_0^t F^0(x_s^*)\,ds + \sum_{j=1}^m \int_0^t F^0\left(\int_0^s U_{s-v}F^j(x_v)\,dw^j(v)\right)ds \\ &\quad + \int_0^t F^0\left(\int_0^s U_{s-v}B^0u(v)\,dv\right)ds \\ &= x^*(t) - y^0 + \sum_{j=1}^m \int_0^t \left(\int_v^t F^0(U_{s-v})\,ds\right)F^j(x_v)\,dw^j(v)\Big) \\ &\quad + \int_0^t \left(\int_v^t F^0(U_{s-v})\,ds\right)B^0u(v)\,dv \\ &= x^*(t) - y^0 + \sum_{j=1}^m \int_0^t (U(t-v)-I)F^j(x_v)\,dw^j(v) \\ &\quad + \int_0^t (U(t-v)-I)B^0u(v)\,dv.\end{aligned}$$

We have used a stochastic version of the Fubini theorem (cf. [I.2]), and the definition of $U(t)$. From this identity we have

$$\bar{x}(t) = y^0 + \int_0^t F^0(\bar{x}_s)\,ds + \sum_{j=1}^m \int_0^t F^j(x_s)\,dw^j(s) + \int_0^t B^0u(s)\,ds,$$

where in particular we see that $\bar{x}(t)$ is a continuous process. Therefore the process $z(t)=\bar{x}(t)-x(t)$ is a continuous process satisfying almost surely the homogeneous equation

$$z(t) = \int_0^t F^0(z_s)\,ds, \qquad t\in[0, T], \quad z_0=0.$$

Since the solution of this equation is unique, we have $z=0$. The proof is complete. □

Remark. The operators D^j are unbounded in Y, and $y(t)$ does not belong to $\mathcal{D}(D^j)$ in general. However, $D^jy(t)$ is well defined by Lemma 2.1. Similarly, $D^jS(t)y$ is well defined as a function in $L^2(0, T; Y)$ for every $y\in Y$, because $D^jS(t)y=(F^j(x_t^*), 0)$, where $x^*(t)$ is the solution of equation (2.3) corresponding to y.

2.2. A basic inequality. In this subsection we prove an inequality for D^j and $S(t)$ which allows us to treat directly the abstract equation (1.2). The direct solution of (1.2) is studied in § 3.

Let us define a new norm in Y by setting

$$|(y^0, y^1)|_\lambda^2 = \|y^0\|^2 + \lambda^2 \int_{-r}^0 \|y^1(s)\|^2\,ds \tag{2.6}$$

for every $(y^0, y^1)\in Y$, where λ is a positive real number. Clearly $|\cdot|_\lambda$ is equivalent to $|\cdot|$, and it is induced by an inner product. We have the following remarkable property.

LEMMA 2.4. *Let D^j and $S(t)$ be defined as in* § 2.1. *Then, for every $c'\in(0, 1)$, there exist $\lambda>0$ and $c''>0$ such that*

$$\sum_{j=1}^m \int_0^{T_1} |D^jS(t)y|_\lambda^2\,dt \leqq (c' + T_1c'')|y|_\lambda^2 \tag{2.7}$$

for every $T_1\in[0, T]$, and $y\in Y$.

Proof. Let $x^*(t)$ be the solution of (2.3) corresponding to a fixed y in Y. By definition of D^j we have

$$\int_0^{T_1} |D^j S(t)y|_\lambda^2 \, dt = \int_0^{T_1} \|F^j(x_t^*)\|^2 \, dt$$

$$\leqq c \int_{-r}^{T_1} \|x^*(t)\|^2 \, dt \quad \text{(by Lemma 2.1)}$$

$$\leqq c \int_{-r}^{0} \|y^1(t)\|^2 \, dt + T_1 \sup_{0 \leqq t \leqq T} \|x^*(t)\|^2$$

$$\leqq c \int_{-r}^{0} \|y^1(t)\|^2 \, dt + cT_1|y|^2$$

(the last inequality is a particular case of (2.2)). Here c denotes a generic positive constant. Thus

$$\sum_{j=1}^{m} \int_0^{T_1} |D^j S(t)y|_\lambda^2 \, dt \leqq mc \int_{-r}^{0} \|y^1(t)\|^2 \, dt + mcT_1|y|^2$$

$$= (mc/\lambda^2)|y|_\lambda^2 + mcT_1|y|_\lambda^2,$$

for $\lambda \geqq 1$. Therefore, given $c' \in (0, 1)$, it is sufficient to choose λ such that $mc/\lambda^2 = c'$ (with $\lambda \geqq 1$) and $c'' = mc$. □

3. Direct solution of the abstract equation (1.2). In this section we show that it is possible to solve directly equation (1.2) in an abstract framework, using some functional analytic properties which hold true for the delay equation (1.1) as well as for the parabolic problem studied in [D.1].

We will assume the following abstract hypotheses:

(A1) Y is a separable Hilbert space; $A: \mathcal{D}(A) \subset Y \to Y$ is the infinitesimal generator of a strongly continuous semigroup $S(t)$ in Y; $D^j: \mathcal{D}(D^j) \subset Y \to Y$ are linear operators, for $j = 1, \cdots, m$;

(A2) there exists a separable Hilbert space V, continuously and densely embedded in Y, with $V \subset \mathcal{D}(D^j)$ and $D^j \in L(V, Y)$ for all $j = 1, \cdots, m$, such that $S(t)V \subset V$ for all $t \geqq 0$ and the restriction of $S(t)$ to V is still strongly continuous;

(A3) there exist two constants $c' \in (0, 1)$ and $c'' > 0$ such that

$$\sum_{j=1}^{m} \int_0^{\tau} |D^j S(t)y|^2 \, dt \leqq (c' + c''\tau)|y|^2 \tag{3.1}$$

for every $y \in V$ and $t \in [0, T]$.

Under these assumptions we study the equation

$$y(t) = S(t)y + \sum_{j=1}^{m} \int_0^t S(t-s)D^j y(s) \, dw^j(s) + \int_0^t S(t-s)f(s) \, ds \tag{3.2}$$

with $y \in Y$ and $f \in M_w^2(0, T; Y)$ (additional terms of the form $\int_0^t S(t-s)f^j(s) \, dw^j(s)$, with $f^j \in M_w^2(0, T; Y)$, can be considered as well).

Remark 1. Taking $V = \mathcal{D}(A)$, we see that assumptions (A1)-(A3) are satisfied by the delay equation (1.1). In particular, the constant c' in (3.1) is even at our choice (see Lemma 2.4).

Remark 2. In the problem studied by [D.1], we can choose $V = \mathcal{D}(A)$ or V equal to the domain of $(-A)^{1/2}$, the fractional power of $(-A)$ with exponent $1/2$. In this case c' is not at our choice, but $c'' = 0$.

Remark 3. In the parabolic case, $S(t)y$ belongs to V for every $t > 0$ and $y \in Y$. This property is no longer true for the delay equation (1.1) (at least for small t).

Since we expect to find solutions of (3.2) in $C_w(0, T; Y)$, the problem of the meaning of the terms $D^j y(\cdot)$ arises. According to Remark 3, and taking $V = \mathcal{D}((-A)^{1/2})$, this problem is overcome in the case of parabolic systems, since we look for solutions in $M_w^2(0, T; V)$. In general we cannot follow this procedure, so that we have to give a special meaning to $D^j y(\cdot)$.

Let us introduce the following classes of processes.

DEFINITION 3.1. Let Φ be the class of all $y(\cdot) \in C_w(0, T; Y)$ that can be represented in the form

$$y(t) = S(t)y + \sum_{j=1}^{m} \int_0^t S(t-s) f^j(s)\, dw^j(s) + \int_0^t S(t-s) f^0(s)\, ds \tag{3.3}$$

for some $y \in Y, f^0, \cdots, f^m \in M_w^2(0, T; Y)$. Similarly, given $y \in Y$ and $f^0 \in M_w^2(0, T; Y)$, let $\Phi(y, f^0)$ be the class of all $y(\cdot) \in C_w(0, T; Y)$ that can be represented in the form (3.3) for some $f^1, \cdots, f^m$ in $M_w^2(0, T; Y)$.

To be unambiguous, this definition requires the following result.

LEMMA 3.1. *If $y(\cdot) \in \Phi$, then the representation* (3.3) *is unique* (*thus the same result holds true in $\Phi(y, f^0)$*).

Proof. It is sufficient to prove the lemma in the case $y(\cdot) = 0$. Since $y = y(0)$, we readily have $y = 0$. Compute $\int_0^t y(s)\, ds$ and apply the stochastic Fubini theorem (cf. [I.2]) to the stochastic integrals of (3.3), and the classical Fubini theorem to the last integral. We obtain

$$\int_0^t y(s)\, ds = A^{-1}\left\{ y(t) - y - \sum_{j=1}^{m} \int_0^t f^j(s)\, dw^j(s) - \int_0^t f^0(s)\, ds \right\},$$

where $\sum_{j=1}^m \int_0^t f^j(s)\, dw^j(s) = -\int_0^t f^0(s)\, ds$ for every $t \in [0, T]$. Since the right-hand side of the last identity is a process of bounded variation, and the left-hand side is a martingale with quadratic variation equal to $\sum_{j=1}^m \int_0^t |f^j(s)|^2\, ds$, it follows that $f^1 = \cdots = f^m = 0$. Therefore, also $f^0 = 0$.

The following lemma allows us to give a meaning to $D^j y(\cdot)$ when $y \in \Phi$.

LEMMA 3.2. (i) *For every $f \in L^2(0, T; V)$ and $\tau \in [0, T]$ we have*

$$\sum_{k=1}^{m} \int_0^\tau \left| D^k \int_0^t S(t-s) f(s)\, ds \right|^2 dt \leqq \tau(c' + c''\tau) \int_0^\tau |f(s)|^2\, ds. \tag{3.4}$$

(ii) *For every $f^1, \cdots, f^m \in M_w^2(0, T; V)$ and $\tau \in [0, T]$ we have*

$$\sum_{k=1}^{m} E \int_0^\tau \left| \sum_{j=1}^{m} D^k \int_0^t S(t-s) f^j(s)\, dw^j(s) \right|^2 dt \leqq (c' + c''\tau) \sum_{j=1}^{m} E \int_0^\tau |f^j(s)|^2\, ds. \tag{3.5}$$

In (3.4) (*respectively*, (3.5)), $\int_0^t S(t-s) f(s)\, ds$ (*respectively*, $\int_0^t S(t-s) f^j(s)\, dw^j(s)$) *is understood as the V-valued integral of a function in $L^2(0, T; V)$* (*respectively, a process in $M_w^2(0, T; V)$*), *well defined because V is a Hilbert space.*

Proof. As to (3.4) we have

$$\sum_{k=1}^{m} \int_0^\tau \left| D^k \int_0^t S(t-s) f(s)\, ds \right|^2 dt = \sum_{k=1}^{m} \int_0^\tau \left| \int_0^t D^k S(t-s) f(s)\, ds \right|^2 dt$$

$$\leqq \sum_{k=1}^{m} \tau \int_0^\tau \int_0^t |D^k S(t-s) f(s)|^2\, ds\, dt$$

$$= \sum_{k=1}^{m} \tau \int_0^\tau \int_s^\tau |D^k S(t-s) f(s)|^2 \, dt \, ds;$$

therefore (3.4) follows from assumption (A3). Since

$$E \left| \sum_{j=1}^{m} \int_0^t D^k S(t-s) f^j(s) \, dw^j(s) \right|^2 = \sum_{j=1}^{m} E \int_0^t |D^k S(t-s) f^j(s)|^2 \, ds,$$

inequality (3.5) can be proved as (3.4).

DEFINITION 3.2. If $y(\cdot) \in \Phi$ (or $y(\cdot) \in \Phi(y, f^0)$) is represented in the form (3.3), then $D^k y(\cdot)$, $k = 1, \cdots, m$, denote the processes in $M_w^2(0, T; Y)$ defined as

$$D^k y(t) = D^k S(t) y + \sum_{j=1}^{m} D^k \int_0^t S(t-s) f^j(s) \, dw^j(s) + D^k \int_0^t S(t-s) f^0(s) \, ds$$

where the terms on the right-hand side are defined by continuity (as elements of $M_w^2(0, T; Y)$) in virtue of (3.1) and Lemma 3.2 (since V and $M_w^2(0, T; V)$ are dense in Y and $M_w^2(0, T; Y)$, respectively).

We can now state the main theorem of this section.

THEOREM 3.3. *Let $y \in Y$ and $f \in M_w^2(0, T; Y)$. Then there exists a unique solution $y(\cdot)$ of (3.4) in the space $\Phi(y, f)$. In particular, $y(\cdot) \in C_w(0, T; Y)$, and $D^j y(\cdot) \in M_w^2(0, T; Y)$ for every $j = 1, \cdots, m$.*

Proof. Consider the system of equations

$$z^k(t) = D^k S(t) y + \sum_{j=1}^{m} D^k \int_0^t S(t-s) z^j(s) \, dw^j(s) + D^k \int_0^t S(t-s) f(s) \, ds, \tag{3.6}$$

$$k = 1, \cdots, m.$$

Let us first show that this system is equivalent to (3.2) via the transformations:

$$z^k(t) = D^k y(t), \qquad k = 1, \cdots, m, \tag{3.7}$$

$$y(t) = S(t) y + \sum_{j=1}^{m} \int_0^t S(t-s) z^j(s) \, dw^j(s) + \int_0^t S(t-s) f(s) \, ds. \tag{3.8}$$

Precisely, if $y(\cdot)$ is a solution of (3.2) in $\Phi(y, f)$, then by definition of $D^k y(\cdot)$ we see that $(z^1(\cdot), \cdots, z^m(\cdot))$ defined by (3.7) belongs to $[M_w^2(0, T; Y)]^m$ and satisfies system (3.6). Conversely, if $(z^1(\cdot), \cdots, z^m(\cdot))$ is a solution of (3.6) in $[M_w^2(0, T; Y)]^m$, and $y(t)$ is defined by (3.8), then $y(\cdot) \in \Phi(y, f)$ and $D^k y(t)$ coincides with $z^k(t)$ for every $k = 1, \cdots, m$, by definition. Then substituting $z^j(s)$ with $D^j y(s)$ into (3.8), we see that $y(t)$ satisfies (3.2). Hence the theorem is proved if system (3.6) has a unique solution in $[M_w^2(0, T; Y)]^m$. But the existence and uniqueness of a local solution of (3.6) can be easily proved by means of inequality (3.5) and the contraction principle in $[M_w^2(0, T_1; Y)]^m$, for sufficiently small T_1. Moreover, the contraction argument can be repeated over intervals of constant length, yielding the global solution. □

Remark. The abstract assumptions of this section do not seem to imply the sample continuity of the solution $y(t)$ of (3.2), in contrast to the concrete approach of § 2.

Using the bound (3.4), along with (3.5), it is also easy to study "closed loop" equations of the form

$$y(t) = S(t) y + \sum_{j=1}^{m} \int_0^t S(t-s) D^j y(s) \, dw^j(s) + \int_0^t S(t-s) G(s) y(s) \, ds \tag{3.9}$$

where $G \in B(0, T; L(Y))$. The same proof as in Theorem 3.3 yields existence and uniqueness of a solution of (3.9) in Φ.

4. Riccati equation and optimal control.

4.1. Statement of the optimal control problem. Let Z be a Hilbert space (the observation space), and $C(t): R^d \times C([-r, 0]; R^d) \to Z$ be a linear operator of the form $C(t)(y^0, y^1) = C^0(t)y^0 + C^1(t)F(y^1)$, where $C^0, C^1 \in B(0, T; L(R^d; Z))$ and F is a bounded linear operator from $C([-r, 0]; R^d)$ into R^d. Let $P_T \in \Sigma^+(Y)$ and $N \in B(0, T; \Sigma^+(R^k))$ such that $N(t) \geqq v$ for every $t \in [0, T]$ and for some constant $v > 0$.

Given $y \in Y$, we consider the problem of minimizing

$$J(u) = E \int_0^T \{|C(t)y(t)|_Z^2 + \|N(t)^{1/2}u(t)\|^2\}\, dt + E|P_T^{1/2}y(T)|^2 \tag{4.1}$$

over all $u \in M_w^2(0, T; R^k)$, where $y(t)$ is the solution of (1.2) given by Theorems 2.3 or 3.3. The process $t \to C(t)y(t)$ is understood as an element of $M_w^2(0, T; Z)$, similarly to the processes $D^k y(t)$ considered in the previous sections.

In the solution of problem (4.1) the central role is played by the Riccati equation (1.3). In § 4.3 we prove the existence of a unique solution P of (1.3) in $B(0, T; \Sigma^+(Y))$ using inequality (2.7). The synthesis of problem (4.1) is solved in § 4.4. The next section is devoted to a preliminary formula which unifies several formulas and equations appearing in the dynamic programming, and will be used in subsequent sections.

4.2. A general identity for the Riccati operator. Denote by $\mathcal{B}$ the class of linear operators $G: R^d \times C([-r, 0]; R^d) \to R^d$ of the form $G(y^0, y^1) = (F(y^1), 0)$, where F is a bounded linear operator from $C([-r, 0]; R^d)$ into R^d. Note that $D^j \in \mathcal{B}$ for every $j = 1, \cdots, m$.

Given $t_0 \in [0, T)$, $x, y \in Y$, $u, v \in M_w^2(t_0, T; R^k)$, and $G^1, \cdots, G^m, H^1, \cdots, H^m \in \mathcal{B}$, consider the following equations on $[t_0, T]$:

$$x(t) = S(t - t_0)x + \sum_{j=1}^m \int_{t_0}^t S(t-s)G^j x(s)\, dw^j(s) + \int_{t_0}^t S(t-s)Bu(s)\, ds, \tag{4.2.i}$$

$$y(t) = S(t - t_0)y + \sum_{j=1}^m \int_{t_0}^t S(t-s)H^j y(s)\, dw^j(s) + \int_{t_0}^t S(t-s)Bv(s)\, ds. \tag{4.2.ii}$$

It is clear that the results of §§ 2 and 3 continue to hold over the interval $[t_0, T]$ instead of $[0, T]$. Thus equations (4.2) have the solutions given by Theorems 2.3 or 3.3.

PROPOSITION 4.1. *Let $P \in B(0, T; \Sigma(Y))$ be a solution of the Riccati equation* (1.3), *and let $x(t)$ and $y(t)$ be the solutions of equations* (4.2.i) *and* (4.2.ii), *respectively. Then*

(4.3)

$$\begin{aligned}
&\langle P(t_0)x, y\rangle \\
&\quad = E\langle P_T x(T), y(T)\rangle \\
&\qquad + E\int_{t_0}^T \sum_{j=1}^m \{\langle P(s)D^j x(s), D^j y(s)\rangle - \langle P(s)G^j x(s), H^j y(s)\rangle\}\, ds \\
&\qquad + E\int_{t_0}^T \{\langle C(s)x(s), C(s)y(s)\rangle_Z + \langle N(s)u(s), v(s)\rangle_k\}\, ds \\
&\qquad - E\int_{t_0}^T \langle N(s)^{-1}(B^*P(s)x(s) + N(s)u(s)), B^*P(s)y(s) + N(s)v(s)\rangle_k\, ds.
\end{aligned}$$

Proof. Let $A_n = nA(n-A)^{-1} \in L(Y)$ be the Yosida approximations of A, defined for sufficiently large $n \in N$ (cf. [Y.1]). Let $S_n(t)$ be the (uniformly continuous) semigroup generated by A_n. We will use the fact that $S_n(t)y$ converges to $S(t)y$ as $n \to \infty$ for every $y \in Y$, uniformly on $[0, T]$. Define $P_n, P_{n,k} \in B(0, T; \Sigma(Y))$ by means of the identities:

$$\begin{aligned} P_n(t)y &= S(T-t)^* P_T S(T-t)y \\ &\quad + \int_t^T S(s-t)^* \left\{ \sum_{j=1}^m D_n^{j*} P(s) D_n^j + C_n(s)^* C_n(s) \right. \\ &\quad \left. - P(s)BN(s)^{-1} BP(s) \right\} S(s-t)y \, ds, \end{aligned} \tag{4.4}$$

$$\begin{aligned} P_{n,k}(t)y &= S_k(T-t)^* P_T S_k(T-t)y \\ &\quad + \int_t^T S_k(s-t)^* \left\{ \sum_{j=1}^m D_n^{j*} P(s) D_n^j + C_n(s)^* C_n(s) \right. \\ &\quad \left. - P(s)BN(s)^{-1} BP(s) \right\} S_k(s-t)y \, ds, \end{aligned}$$

$t \in [0, T]$, $y \in Y$, where $D_n^j = D^j n(n-A)^{-1}$ and $C_n(t) = C(t)n(n-A)^{-1}$.

From the convergence results proved in the Appendix we have $\lim_{n\to\infty} \langle P_n(t)x, y\rangle = \langle P(t)x, y\rangle$ for every $x, y \in Y$ and $t \in [0, T]$. Moreover, from the convergence property of $S_k(t)$, we have $\lim_{k\to\infty} \langle P_{n,k}(t)x, y\rangle = \langle P_n(t)x, y\rangle$ for every $n \in N$, $x, y \in Y$, and $t \in [0, T]$. Let us also approximate $x(t)$ and $y(t)$ by the solutions $x_n(t)$ and $y_n(t)$ of equations (2.4.i, ii) with G_n^j and H_n^j in place of G^j and H^j, where $G_n^j = G^j n(n-A)^{-1}$ and $H_n^j = H^j n(n-A)^{-1}$. Moreover let $x_{n,k}(t)$ be the unique solution of the equation

$$x_{n,k}(t) = S_k(t-t_0)x + \sum_{j=1}^m \int_{t_0}^t S_k(t-s) G_n^j x_{n,k}(s) \, dw^j(s) + \int_{t_0}^t S_k(t-s)Bu(s) \, ds,$$

and let $y_{n,k}(t)$ be similarly defined.

Since the operators in (4.4) are bounded, we can differentiate (4.4). After some computations we have

$$\begin{aligned} &dP_{n,k}(t)y/dt + A_k^* P_{n,k}(t)y + P_{n,k}(t)A_k y \\ &\quad + \left\{ \sum_{j=1}^m D_n^{j*} P(t) D_n^j + C_n(t)^* C_n(t) - P(t)BN(t)^{-1} B^* P(t) \right\} y = 0 \end{aligned}$$

for every $t \in [0, T]$ and $y \in Y$. Applying the Ito formula (cf. [I.2]) to $\langle P_{n,k}(t)x_{n,k}(t), y_{n,k}(t)\rangle$ over $[t_0, T]$, after some manipulations we obtain

$$\begin{aligned} &E\langle P_T x_{n,k}(T), y_{n,k}(T)\rangle \\ &= \langle P_{n,k}(t_0)x, y\rangle + \sum_{j=1}^m E \int_{t_0}^T \langle (H_n^{j*} P_{n,k}(s) G_n^j - D_n^{j*} P(s) D_n^j) x_{n,k}(s), y_{n,k}(s)\rangle \, ds \\ &\quad - E \int_{t_0}^T \{\langle C_n(s)x_{n,k}(s), C_n(s)y_{n,k}(s)\rangle_Z + \langle N(s)u(s), v(s)\rangle_k\} \, ds \\ &\quad + E \int_{t_0}^T \{\langle N(s)^{-1} B^* P(s)x_{n,k}(s), B^* P(s)y_{n,k}(s)\rangle_k + \langle u(s), B^* P_{n,k}(s)y_{n,k}(s)\rangle_k \\ &\qquad + \langle B^* P_{n,k}(s)x_{n,k}(s), v(s)\rangle_k + \langle N(s)u(s), v(s)\rangle_k\} \, ds. \end{aligned}$$

Since $\lim_{k\to\infty} \sup_{t_0\le t\le T} E|x_{n,k}(t)-x_n(t)|^2=0$ for every $n\in N$ (this result is classical because the operators in the equations for $x_{n,k}$ and x_n are bounded; see [I.2]), and the same result holds true for $y_{n,k}$ and y_n, we can first take the limit as $k\to\infty$ in the last identity. Then we can take the limit as $n\to\infty$, using the convergence results proved in the Appendix. By these limit procedures we obtain the desired result. □

We conclude this section by proving a partial converse of Proposition 4.1. Consider equations (4.2) with $u=v=0$, and $G^j=H^j=D^j$. Denote by $x(t, t_0; x)$ the solution of (4.2.i). Since $x(t, t_0; x)=x(t-t_0; 0; x)$, we simply denote it by $x(t-t_0; x)$. The process $y(t-t_0; y)$ is similarly defined. To summarize, we have

$$x(t-t_0; x)=S(t-t_0)x+\sum_{j=1}^{m}\int_{t_0}^{t} S(t-s)D^j x(s-t_0; x)\,dw^j(s), \tag{4.5.i}$$

$$y(t-t_0; y)=S(t-t_0)y+\sum_{j=1}^{m}\int_{t_0}^{t} S(t-s)D^j y(s-t_0; y)\,dw^j(s). \tag{4.5.ii}$$

PROPOSITION 4.2. *$P\in B(0, T; \Sigma(Y))$ is a solution of the Riccati equation* (1.3) *if and only if it is a solution of the equation*

$$\begin{aligned}\langle P(t)x, y\rangle &= E\langle P_T x(T-t; x), y(T-t; y)\rangle \\ &\quad + E\int_t^T \langle C(s)x(s-t; x), C(s)y(s-t; y)\rangle_Z\,ds \\ &\quad - E\int_t^T \langle N(s)^{-1}B^*P(s)x(s-t; x), B^*P(s)y(s-t; y)\rangle_k\,ds\end{aligned} \tag{4.6}$$

for every $t\in[0, T]$ and $x, y\in Y$, where $x(t; x)$ and $y(t; y)$ are defined by equations (4.5).

Proof. It is sufficient to prove that (4.6) implies (1.3). Let $P(t)$ be a solution of (4.6), and let $P_n(t)$ be the unique solution in $B(0, T; \Sigma(Y))$ of the linear equation

$$\begin{aligned}\langle P_n(t)x, y\rangle &= \langle P_T S(T-t)x, S(T-t)y\rangle + \sum_{j=1}^{m}\int_t^T \langle P_n(s)D_n^j S(s-t)x, D_n^j S(s-t)y\rangle\,ds \\ &\quad + \int_t^T \langle\{C_n(s)^*C_n(s)-P(s)BN(s)^{-1}B^*P(s)\}S(s-t)x, S(s-t)y\rangle\,ds\end{aligned} \tag{4.7}$$

$x, y\in Y$, $t\in[0, T]$. Here the operators D_n^j and $C_n(s)$ are defined as in the proof of the previous proposition. This equation is covered by [I.1], because D_n^j and $C_n(s)$ are bounded operators in Y ((4.7) can be studied by standard contraction arguments). In [I.1] it is also proved that

$$\begin{aligned}\langle P_n(t)x, y\rangle &= E\langle P_T x_n(T-t; x), y_n(T-t; y)\rangle \\ &\quad + E\int_t^T \langle\{C_n(s)^*C_n(s)-P(s)BN(s)^{-1}B^*P(s)\}x_n(s-t; x), y_n(s-t; y)\rangle\,ds\end{aligned} \tag{4.8}$$

where $x_n(t; x)$ and $y_n(t; y)$ are the approximations of $x(t; x)$ and $y(t; y)$, defined in the previous proof. Note that (4.8) can also be obtained by Proposition 4.1. From (4.8) and the convergence results proved in the Appendix it follows that $\langle P_n(t)x, y\rangle\to\langle P(t)x, y\rangle$ as $n\to\infty$, for every $t\in[0, T]$ and $x, y\in Y$. Thus it is sufficient to take the limit as $n\to\infty$ in (4.7) to see that P satisfies (1.3). □

4.3. Direct solution of (1.3).

THEOREM 4.3. *There exists a unique solution $P\in B(0, T; \Sigma^+(Y))$ of the Riccati equation* (1.3).

Proof. Let us begin by changing the metric underlying (1.3). Let $\lambda > 0$ be given by Lemma 3.1 with respect to some $c' \in (0, 1)$. The norm $|\cdot|_\lambda$ is induced by the inner product $\langle x, y\rangle_\lambda = \langle x^0, y^0\rangle_d + \lambda^2 \int_{-r}^0 \langle x^1(s)y^1(s)\rangle_d\, ds$, where $x = (x^0, x^1)$ and $y = (y^0, y^1)$ are elements of Y. Therefore there exists an invertible operator $Q \in L(Y)$ such that $\langle x, y\rangle = \langle Qx, y\rangle_\lambda = \langle x, Qy\rangle_\lambda$ (it is defined as $Q(y^0, y^1) = (y^0, \lambda^2 y^1)$). If $P(t)$ is a solution of (1.3) in $B(0, T; L(Y))$, then $\tilde{P}(t) = QP(t)$ is a solution in $B(0, T; L(Y))$ of the Riccati equation

(4.9)

$$\begin{aligned}\langle \tilde{P}(t)x, y\rangle_\lambda = {} & \langle \tilde{P}_T S(T-t)x, S(T-t)y\rangle_\lambda \\ & + \sum_{j=1}^m \int_t^T \langle \tilde{P}(s)D^j S(s-t)x, D^j S(s-t)y\rangle_\lambda\, ds \\ & + \int_t^T \langle C(s)S(s-t)x, C(s)S(s-t)y\rangle_Z\, ds \\ & - \int_t^T \langle N(s)^{-1}\tilde{B}^*\tilde{P}(s)S(s-t)x, \tilde{B}^*\tilde{P}(s)S(s-t)y\rangle_k\, ds, \quad x, y \in Y,\end{aligned}$$

where $\tilde{P}_T = QP_T$, $\tilde{B}^* = B^*Q^{-1}$. Since Q is invertible, the converse is also true.

Due to the inequality (2.7), it is easy to see that the contraction principle in $B(T_1, T; L(Y))$ can be applied to equation (4.9), as in [D-I.1], for $T - T_1$ sufficiently small. Therefore, by the equivalence between (1.3) and (4.9), there exists a unique solution P of (1.3) in $B(T_1, T; L(Y))$. Since $P_T \in \Sigma^+(Y)$, we see from (1.3) that $P(t) \in \Sigma(Y)$ for every $t \in [T_1, T]$ (the same argument could be applied to $\tilde{P}(t)$, because it is possible to show that also $\tilde{P}_T \in \Sigma^+(Y)$ with respect to $\langle\cdot,\cdot\rangle_\lambda$). Iterating this procedure we get a unique maximal solution of (1.3), such that $P(t) \in \Sigma(Y)$ for every t in the maximal interval of existence. To show that $P(\cdot)$ is in fact a global solution, we prove an a priori bound for $P(\cdot)$. Let us apply Proposition 4.1 to $P(t)$, with an arbitrary t_0, but choosing $x = y$, $u = v$, and $G^j = H^j = D^j$. Thus

$$\langle P(t_0)y, y\rangle = J_{t_0}(u) - E\int_{t_0}^T \|N(s)^{-1/2}B^*P(s)y(s) + N(s)^{1/2}u(s)\|^2\, ds, \tag{4.10}$$

where

$$J_{t_0}(u) = E\int_{t_0}^T \{|C(s)y(s)|_Z^2 + \|N(s)^{1/2}u(s)\|^2\}\, ds + E|P_T^{1/2}y(T)|^2.$$

If we take $u = 0$ in (4.10) we can find a constant $c > 0$ independent of $y \in Y$ and $t_0 \in [0, T]$, such that

$$\langle P(t_0)y, y\rangle \leqq c|y|^2. \tag{4.11}$$

On the other hand, if we choose

$$u(t) = -N(t)^{-1}B^*P(t)y(t) \tag{4.12}$$

in (4.10), we have $\langle P(t_0)y, y\rangle = J_{t_0}(u) \geqq 0$. Therefore (4.11) yields $|P(t)| \leqq c$ for every t in the maximal interval of existence, which is the required a priori bound. We have also proved that $P(t) \in \Sigma^+(Y)$.

Remark 1. We meet directly (4.9) if we rewrite the cost functional in (4.1) using the inner product $\langle\cdot,\cdot\rangle_\lambda$:

$$J(u) = E\int_0^T \{|C(s)y(s)|_Z^2 + \|N(s)^{1/2}u(s)\|^2\}\, ds + \langle \tilde{P}_T y(T), y(T)\rangle_\lambda,$$

where $\tilde{P}_T = QP_T$, Q defined in the proof of the previous theorem. As to this, notice that $\tilde{B}^*$ is just the adjoint of B with respect to $\langle \cdot, \cdot \rangle_\lambda$.

Remark 2. The Riccati operator $P(t)$ can also be obtained by the method of characteristics introduced by [D.2]. This method consists of the direct solution of (4.6), which is equivalent to (1.3) by Proposition 4.2. The existence and uniqueness of a local (and maximal) solution of (4.6) can be proved by the contraction principle, while the global existence follows from an a priori bound similar to the one proved in Theorem 4.3.

4.4. Synthesis. Finally, we have the following theorem.

THEOREM 4.4. *For every fixed $y \in Y$, there exists a unique optimal control u^* for problem* (4.1) *in $M_w^2(0, T; R^k)$, characterized by the feedback formula*

$$u^*(t) = -N(t)^{-1} B^* P(t) y^*(t), \qquad 0 \leqq t \leqq T, \tag{4.13}$$

where $P(t)$ is the unique solution in $B(0, T; \Sigma^+(Y))$ of (1.3), *and $y^*(t)$ is the solution of* (1.2) *corresponding to $u^*(t)$. In other words, $y^*(t) = (x^*(t), x_t^*)$, where $x^*(t)$ is the unique solution of the closed loop equation*

$$\begin{aligned} x^*(t) = y^0 &+ \int_0^t F^0(x_s^*)\, ds + \sum_{j=1}^m \int_0^t F^j(x_s^*)\, dw^j(s) \\ &- \int_0^t B^0 N(s)^{-1} B^* P(s)(x^*(s), x_s^*)\, ds, \quad x_0^* = y^1. \end{aligned} \tag{4.14}$$

Finally, $J(u^*) = \langle P(0)y, y \rangle$.

Proof. It is sufficient to take $t_0 = 0$ in (4.10); it follows that $\langle P(0)y, y \rangle \leqq J(u)$ for every $u \in M_w^2(0, T; R^k)$, and $\langle P(0)y, y \rangle = J(u^*)$ if and only if u^* satisfies (4.13). □

Appendix. Throughout this Appendix we shall assume the hypotheses of § 2. Let D be an operator in the class $\mathcal{B}$ defined in § 4.2. Clearly $D \in L(\mathcal{D}(A), Y)$, and there exists a constant $c > 0$ such that

$$\int_0^T |DS(t)y|^2\, dt \leqq c|y|^2 \quad \forall y \in Y \tag{A.1}$$

(this can be proved as inequality (2.7)).

PROPOSITION. *Let D be given as above, and let $D_n = DI_n$, where $I_n = n(n - A)^{-1}$ (for $n \in N$ large enough). Then*

$$D_n S(\cdot)y \to DS(\cdot)y \text{ in } L^2(0, T; Y) \text{ as } n \to \infty, \quad \forall y \in Y. \tag{A.2}$$

Moreover, let $y(\cdot)$ be the solution of (3.2), *and let $y_n(\cdot)$ be the solution of the approximating equation*

$$y_n(t) = S(t)y + \sum_{j=1}^m \int_0^t S(t-s) D_n^j y_n(s)\, dw^j(s) + \int_0^t S(t-s) f(s)\, ds \tag{A.3}$$

where $D_n^j = D^j I_n$. Then

$$y_n(\cdot) \to y(\cdot) \text{ in } C_w(0, T; Y) \text{ and } D_n y_n(\cdot) \to Dy(\cdot) \text{ in } M_w^2(0, T; Y) \tag{A.4}$$

as $n \to \infty$.

Proof. Since $D_n S(t)y - DS(t)y = DS(t)[I_n y - y]$, (A.2) readily follows from (A.1). Similarly we can prove that

$$D_n^j y(\cdot) \to D^j y(\cdot) \text{ and } D_n y(\cdot) \to Dy(\cdot) \text{ as } n \to \infty \text{ in } M_w^2(0, T; Y). \tag{A.5}$$

Let us now prove the first part of (A.4). Let $z_n(t) = y_n(t) - y(t)$. Then

$$z_n(t) = \sum_{j=1}^{m} \int_0^t S(t-s)\{D_n^j z_n(s) + [D_n^j - D^j]y(s)\}\, dw^j(s). \tag{A.6}$$

Given $c' \in (0, 1)$, let c'' and λ be given by Lemma 2.4. Then, from Lemma 3.2,

$$\begin{aligned}
&\sum_{k=1}^{m} E \int_0^\tau |D_n^k z_n(t)|_\lambda^2\, dt \\
&\quad \leqq (c' + c''\tau) \sum_{j=1}^{m} E \int_0^\tau |I_n\{D_n^j z_n(s) + [D_n^j - D^j]y(s)\}|_\lambda^2\, ds \\
&\quad \leqq (c' + c''\tau)2c \sum_{j=1}^{m} E \int_0^\tau |D_n^j z_n(s)|_\lambda^2\, ds + (c' + c''\tau)2c \sum_{j=1}^{m} E \int_0^\tau |[D_n^j - D^j]y(s)|_\lambda^2\, ds
\end{aligned}$$

where c is a constant greater than $|I_n|^2$ for every n (it exists by the Hille-Yosida theorem). Thus, taking c' such that $2cc' < 1$ and then τ such that $2c(c' + c''\tau) < 1$, we see that

$$\sum_{k=1}^{m} E \int_0^\tau |D_n^k z_n(t)|_\lambda^2\, dt \to 0 \quad \text{as } n \to \infty \tag{A.7}$$

in virtue of the convergence result (A.5). This argument can be repeated over intervals of constant length, so that (A.7) holds true with $\tau = T$. Now using (A.5) and (A.7) in (A.6) we obtain the first convergence result stated in (A.4). As to the second one, from (A.6) we have

$$D_n z_n(t) = \sum_{j=1}^{m} D_n \int_0^t S(t-s) I_n\{D_n^j z_n(s) + [D_n^j - D^j]y(s)\}\, dw^j(s). \tag{A.8}$$

Thus, again using (A.5) and (A.7) in (A.8), we see that $D_n z_n(\cdot) \to 0$ in $M_w^2(0, T; Y)$ (note that inequalities of the form (3.4) and (3.5) hold true for D in place of D^k, by virtue of (A.1), replacing the constant $(c' + c''\tau)$ of (3.4) and (3.5) with the constant c of (A.1)). Therefore, $D_n y_n(\cdot) - Dy(\cdot) = D_n z_n(\cdot) - [D_n - D]y(\cdot) \to 0$ in $M_w^2(0, T; Y)$ (recall (A.5)), completing the proof. □

REFERENCES

[D.1] G. DA PRATO, *Some results on linear stochastic evolution equations in Hilbert spaces by the semigroup method*, Stochastic Anal. Appl., 1 (1983), pp. 57-88.

[D.2] ———, *Direct solution of a Riccati equation arising in stochastic control theory*, Appl. Math. Optim., 11 (1984), pp. 191-208.

[D-I.1] G. DA PRATO and A. ICHIKAWA, *Stability and quadratic control for linear stochastic equations with unbounded coefficients*, Scuola Normale Superiore, 1984, preprint.

[D-M.1] M. C. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constant delays. 1. General case*, J. Differential Equations, 12 (1972), pp. 213-235.

[D-S.1] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* I: *General Theory*, Wiley Interscience, New York, 1967.

[F.1] F. FLANDOLI, *Direct solution of a Riccati equation arising in a stochastic control problem with control and observation on the boundary*, Appl. Math. Optim., 14 (1986), pp. 107-129.

[I.1] A. ICHIKAWA, *Dynamic programming approach to stochastic evolution equations*, SIAM J. Control Optim., 17 (1979), pp. 152-173.

[I.2] ———, *Stability of semilinear stochastic evolution equations*, J. Math. Anal. Appl., 90 (1982), pp. 12-44.

[M.1] S. MOHAMMED, *Stochastic Functional Differential Equations*, Research Notes in Math. 99, Pitman, London, 1984.

[Y.1] K. YOSIDA, *Functional Analysis*, 123, Springer-Verlag, Berlin, 1965.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 950-965, July 1990

010

AN APPROXIMATION SCHEME FOR THE MINIMUM TIME FUNCTION*

M. BARDI† AND M. FALCONE‡

Abstract. This paper presents an approximation scheme for the nonlinear minimum time problem with compact target. The scheme is derived from a discrete dynamic programming principle and the main convergence result is obtained by applying techniques related to discontinuous viscosity solutions for Hamilton-Jacobi equations. The convergence is proved under general controllability assumptions on both the continuous-time and the discrete-time systems. An explicit sufficient condition on the system and the target ensuring the desired controllability is given. This condition is shown to be necessary and sufficient for the Lipschitz continuity of the minimum time function if the target is smooth. An extension to the case of a point-shaped target is given.

Key words. discrete dynamic programming, viscosity solution weak limits, Hamilton-Jacobi equation, time-optimal control

AMS(MOS) subject classifications. 49C20, 93C55

1. Introduction. In this paper we begin a study of the minimum time problem from a computational point of view in the framework of dynamic programming.

We consider a controlled dynamical system

$$\begin{cases} y' = b(y, \alpha) \\ y(0) = x \end{cases} \tag{1.1}$$

$y \in \mathbb{R}^N$, $\alpha(t) \in A \subseteq \mathbb{R}^M$, and a compact target set $\mathcal{T}$. We call $\mathcal{R}$ the set of all initial points from which the system can reach $\mathcal{T}$ in finite time, and $T(x)$, $x \in \mathcal{R}$, the infimum of the times necessary to reach $\mathcal{T}$ starting at x. If $\mathcal{T} = \{0\}$, then T is the classical minimum time function, see e.g., Lee and Markus [25], Hermes and La Salle [21], Conti [11], and Bacciotti [1]. The dynamic programming principle (see e.g., [18]) leads to associating to this problem the Hamilton-Jacobi-Bellman partial differential equation

$$\sup_{a \in A} \{-b(x, a) \cdot \nabla T(x)\} = 1 \quad \text{in } \mathcal{R} \setminus \mathcal{T} \tag{1.2}$$

(where $\cdot$ stands for the scalar product). P. L. Lions [26] has shown that the value functions of a large class of deterministic control problems satisfy the dynamic programming equation in the *viscosity* sense, a concept introduced by Crandall and Lions [13]. As far as the minimum time problem is concerned, Bardi [3] has recently proved that, if the system is locally controllable around the whole target $\mathcal{T}$, then T is the unique viscosity solution of (1.2) satisfying the boundary conditions

$$\begin{aligned} &T(x) = 0 \qquad \text{on } \partial\mathcal{T} \\ &T(x) \to +\infty \quad \text{as } x \to \partial\mathcal{R}. \end{aligned} \tag{1.3}$$

A basic role in that proof is played by the change of unknown variable

$$v(x) = 1 - e^{-T(x)} \tag{1.4}$$

* Received by the editors February 10, 1989; accepted for publication (in revised form) October 10, 1989. Most of the research for this paper was done while the authors visited the Centre de la Recherche de Mathematique de la Décision, Paris.

† Dipartimento di Matematica Pura e Applicata, Universita' di Padova, via Belzoni 7, 35131 Padova, Italy. The work of this author was partially supported by the Italian Consiglio Nazionale delle Ricerche.

‡ Dipartimento di Matematica, Universita' di Roma "La Sapienza," P. Aldo Moro 2, 00185 Roma, Italy.

first used by Kruzkov [24], which transforms (1.2), (1.3) into

$$\begin{cases} v(x)+\sup_{a\in A}\{-b(x,a)\cdot\nabla v(x)\}=1 & \text{in } \mathcal{R}\setminus\mathcal{T} \\ v=0 & \text{on } \partial\mathcal{T} \\ v=1 & \text{on } \partial\mathcal{R}. \end{cases} \tag{1.5}$$

In this paper we change slightly the point of view in that we consider $\mathcal{R}$ as an unknown set. We remark that, since v is itself the value function of a control problem with discount rate, it satisfies also the boundary value problem

$$\begin{cases} v(x)+\sup_{a\in A}\{-b(x,a)\cdot\nabla v(x)\}=1 & \text{in } \mathbb{R}^N\setminus\mathcal{T} \\ v=0 & \text{on } \partial\mathcal{T}. \end{cases} \tag{1.6}$$

Once this problem is solved we immediately recover both T and $\mathcal{R}$ via the formulas

$$T(x)=-\log(1-v), \qquad \mathcal{R}=\{x\colon v(x)<1\}. \tag{1.7}$$

Therefore, we will look for a numerical approximation of (1.6) and we will employ to this end the notion of viscosity solution. The viscosity solution approach has already been used to prove the convergence of discrete approximation schemes for Hamilton-Jacobi equations and for control problems by Capuzzo Dolcetta [8], Capuzzo Dolcetta and Ishii [10], Souganidis [30], Falcone [17], (see also the survey paper of Capuzzo Dolcetta and Falcone [9] for a more extensive list of references). As in [8] and [10], we approximate (1.1) via a one-step scheme (for simplicity we adopt a Euler scheme of step h) and consider the corresponding discrete-time control problem which consists of minimizing the number of steps necessary to reach the target starting at x. Let us denote $N_h(x)$ this minimum number of steps and define

$$v_h(x)=1-e^{-hN_h(x)}. \tag{1.8}$$

The dynamic programming equation for $v_h(x)$ is the following discretized version of (1.6):

$$\begin{cases} v_h(x)+\sup_{a\in A}\{-e^{-h}v_h(x+hb(x,a))\}=1-e^{-h} & \text{in } \mathbb{R}^N\setminus\mathcal{T} \\ v_h=0 & \text{on } \partial\mathcal{T}, \end{cases} \tag{1.9}$$

and it is equivalent to a fixed point problem for a contraction mapping in L^∞. Once (1.9) is solved, dynamic programming provides an optimal control in feedback form for the discrete-time problem. However, by its very definition, v_h is piecewise constant and discontinuous, so that the standard compactness argument based on the Ascoli-Arzela theorem used in [8] to pass to the limit as $h\to 0$ does not apply here. To overcome this difficulty and show, nonetheless, the uniform convergence $v_h\to v$, we introduce a method inspired by some recent strong stability results for Hamilton-Jacobi equations proved by Barles and Perthame [5] in the framework of the theory of discontinuous viscosity solutions (see also [6] and Ishii's papers [22], [23]). The main idea of this method is to take "weak limits" based on just L^∞ estimates and then use a strong comparison result between semicontinuous sub- and super-solutions of the limit equation (1.6). This works if both the continuous and the discrete systems are locally controllable around $\mathcal{T}$, in the sense that (i) T is continuous, and (ii) the following condition holds for the discrete system:

$$v_h(x)\leqq\sigma(\operatorname{dist}(x,\mathcal{T}),h) \tag{1.10}$$

for x sufficiently close to $\mathcal{T}$ and h sufficiently small, where σ is continuous and $\sigma(0,0)=0$.

In view of the recent deep developments of "geometric control theory" (see, e.g., [1], [2], [11], [28], [31] and the references therein) our results should be applicable in a variety of interesting cases. However, most of the literature in this field is devoted to studying controllability for the continuous system around a point-shaped target, so it is not directly applicable here (in general it is necessary to assume the target has nonempty interior in order to hope to hit it with the discrete system, that is to have (1.10)). Therefore we look for sufficient conditions on the system ensuring the desired controllability in the sense of (i) and (ii), for targets with piecewise smooth boundary. We limit ourselves here to a "zeroth order condition," namely a condition just on the fields b and not on their derivatives (if they exist), their Lie brackets, and so on. Roughly speaking, the sufficient condition is:

(1.11) in each point of $\partial\mathcal{T}$ there is a vector field b pointing inward $\mathcal{T}$.

We show that (1.11) is necessary and sufficient for the local Lipschitz continuity of T if the target is smooth, so that it plays the same role for such targets as Petrov's positive basis condition does for the case $\mathcal{T}=\{0\}$ (see [28], [29], [33]). For C^2 targets, Friedman [32] showed the sufficiency of condition (1.11) for the Lipschitz continuity of T, in the more general context of differential games.

In a forthcoming paper [4] we will establish estimates of the rate of convergence of the above discrete approximation.

Other authors have recently studied the minimum time problem using the theory of viscosity solutions. A local uniqueness result for equation (1.2) has been proved by Evans and James [15], while Hermes [20] has proposed a method for studying the local structure of optimal feedback controls employing the Bellman equation. Combining the change of variables (1.4) with the methods of Barles and Perthame [5] it is also possible to give a uniqueness result for semicontinuous solutions of (1.2), plus a modified boundary condition on $\partial\mathcal{T}$ (see Remark 3.5 below) provided the target is smooth. This is interesting for problems lacking controllability on some parts of $\partial\mathcal{T}$.

Other numerical methods for the minimum time problem, mostly for the linear case with point-shaped target, can be found in Neustadt [27], Eaton [14], Fujisawa and Yasuda [19], Canon, Cullum, and Polak [7] and Falb and de Jong [15]. We refer the interested reader to the Appendix where we discuss the main differences between these methods and ours.

The paper is organized as follows. In § 2 we lay down the hypotheses and study the discrete-time problem. In § 3 we prove the general convergence result as $h\to 0$. In § 4 we show that (1.11) implies the local controllability of the discrete system. In § 5 the same is done for the continuous system, the Lipschitz continuity of T is studied, and a final remark shows a way to apply these techniques also to the classical case $\mathcal{T}=\{0\}$. The Appendix, § 6, compares our method to other approximation schemes.

2. The minimum time problem and its discretization. Before giving the discrete version of the minimum time problem we recall, for the reader's convenience, its continuous version and some assumptions and results which we will use in the sequel of this paper. We assume that the *set of admissible controls* A is a subset of $\mathbb{R}^M$ and we define the set $\mathcal{A}$ of *admissible control functions* appearing in (1.1) to be

$$\mathcal{A}\equiv\{\alpha:[0,+\infty[\to A,\ \text{measurable}\}.$$

Let us denote $y_x(t,\alpha)\equiv y_x(t)$ the solution of (1.1) and define

$$t_x(\alpha)=\inf\{t:\ y_x(t)\in\mathcal{T}\}\leqq+\infty, \tag{2.1}$$

where $t_x = +\infty$ if $y_x(\cdot)$ never reaches the target. As we said in the introduction, our problem consists in minimizing the time necessary to reach the target $\mathcal{T}$, that is, to obtain

$$T(x) \equiv \inf_{\alpha \in \mathcal{A}} t_x(\alpha). \tag{2.2}$$

We define the set

$$\begin{aligned} \mathcal{R} &\equiv \{x \in \mathbb{R}^N : T(x) < +\infty\} \\ &= \{x \in \mathbb{R}^N : \text{there exists } \alpha \in \mathcal{A} \text{ and } t \geqq 0 \text{ such that } y_x(t) \in \mathcal{T}\}. \end{aligned} \tag{2.3}$$

Obviously $\mathcal{R} \supseteq \mathcal{T}$ and $T(x) = 0$ for any $x \in \mathcal{T}$. We will use the following assumptions

$$\text{(A1)} \quad \begin{cases} b : \mathbb{R}^N \times A \to \mathbb{R}^N \text{ is continuous;} \\ |b(x, a) - b(y, a)| \leqq L|x - y| \text{ and } |b(y, a)| \leqq K(1 + |y|) \\ \text{for all } x, y \in \mathbb{R}^N \text{ and } a \in A; \\ \mathcal{T} \text{ is compact.} \end{cases}$$

The next theorem gives the connection between the control problem and a well-posed boundary value problem for a Hamilton-Jacobi partial differential equation (PDE), whose solution determines both T and $\mathcal{R}$.

THEOREM 2.1. *Assume* (A1), *T continuous, and define*

$$v(x) \equiv \begin{cases} 1 - e^{-T(x)} & \text{for } x \in \mathcal{R}; \\ 1 & \text{for } x \notin \mathcal{R}. \end{cases}$$

Then v is the unique bounded viscosity solution of (1.6).

Proof. We can write

$$v(x) = \inf_{\alpha \in \mathcal{A}} \int_0^{t_x(\alpha)} e^{-t}\, dt,$$

which shows that v is the value function of a control problem "with interest rate." Then the Dynamic Programming Principle implies that v is a viscosity solution of (1.6), by the arguments of [26]. The uniqueness of bounded continuous solutions of (1.6) is proved in [3]. □

We can obtain a discrete version of the minimum time problem in the following way. Let us first choose a step in time h, $h > 0$, and define the sequence

$$t_j \equiv jh. \tag{2.4}$$

We shall assume that the state is observed only at discrete times t_j so we replace (1.1) by the recursive sequence

$$\begin{cases} x_{j+1} = x_j + hb(x_j, a_j) \\ x_0 = x \end{cases} \tag{2.5}$$

where $x_j = x_{t_j}$, $a_j = a_{t_j}$, and $a_j \in A$ (just to simplify notation, x_j and a_j will sometimes also denote the whole sequences $\{x_j\}$ and $\{a_j\}$). For any given x and $\{a_j\}$, we define the function

$$n_h(a_j, x) \equiv \min\{j \in \mathbb{N} : x_j \in \mathcal{T}\} \leqq +\infty, \tag{2.6}$$

where $n_h = +\infty$ if x_j never reaches the target $\mathcal{T}$. We want to determine, for any x, the minimum number of steps necessary to reach $\mathcal{T}$, that is

$$N_h(x) \equiv \min_{\{a_j\}} n_h(a_j, x). \tag{2.7}$$

We define the set

$$\begin{aligned}\mathcal{R}_h &\equiv \{x \in \mathbb{R}^N : N_h(x) < +\infty\} \\ &= \{x \in \mathbb{R}^N : \text{there exists } \{a_j\} \text{ and } j \in \mathbb{N} \text{ such that } x_j \in \mathcal{T}\}.\end{aligned} \tag{2.8}$$

As in the continuous version we will make use of the change of variable (1.4) and define the cost

$$J_x^h(\{a_j\}) \equiv 1 - e^{-hn_h(a_j,x)} = \left[\sum_{j=0}^{n_h(a_j,x)-1} e^{-jh}\right](1-e^{-h})\chi_{C\mathcal{T}}(x) \tag{2.9}$$

where $\chi_{C\mathcal{T}}(x)$ is the characteristic function of $C\mathcal{T} = \mathbb{R}^N \setminus \mathcal{T}$. Then the value function for this problem will be given by

$$v_h(x) = \begin{cases} \min\limits_{\{a_j\}} J_x^h(\{a_j\}) = 1 - e^{-hN_h(x)} & \text{for any } x \in \mathcal{R}_h \\ 1 & \text{for any } x \notin \mathcal{R}_h. \end{cases} \tag{2.10}$$

The following proposition is well known.

PROPOSITION 2.2. (Discrete Dynamic Programming Principle). *The value function* v_h *verifies*

$$v_h(x) = \min_{\{a_j\}} \left[(1-\beta)\sum_{p=0}^{q-1} \beta^p + v_h(x_q)\beta^q\right], \quad \textit{where } \beta \equiv e^{-h}, \tag{DDPP}$$

for any x *belonging to* $\mathcal{R}_h \setminus \mathcal{T}$ *and* $0 < q \leqq N_h(x)$.

Proof. From (2.9) and (2.10), there exists a sequence $\{a_j^*\}$ such that

$$v_h(x) = J_x^h(\{a_j^*\}) \geqq (1-\beta)\sum_{p=0}^{q-1} \beta^p + v_h(x_q)\beta^q$$

for $0 < q \leqq N_h(x)$, and we obtain the first inequality. The converse inequality is easily seen by the definition of v_h. □

The (DDPP) for $q = 1$ and $x \in \mathcal{R}_h \setminus \mathcal{T}$ gives

$$v_h(x) = \min_{a \in A} [\beta v_h(x + hb(x, a))] + (1-\beta) \tag{HJ_h}$$

which, due to the definition of $\mathcal{R}_h$ and (2.10), is valid not only in $\mathcal{R}_h \setminus \mathcal{T}$ but in

$$\Omega \equiv \mathbb{R}^N \setminus \mathcal{T}.$$

THEOREM 2.3. *The function* v_h *is the unique bounded solution of* (1.9), *i.e., of*

$$\begin{cases} u(x) = Su(x) & \forall x \in \Omega \\ u(x) = 0 & \forall x \in \mathcal{T} \end{cases} \tag{2.11}$$

where

$$Su(x) \equiv \inf_{a \in A} \{\beta u(x + hb(x, a))\} + 1 - \beta. \tag{2.12}$$

Proof. Let us observe that if $u : \mathbb{R}^N \to \mathbb{R}$ is bounded, then $Su > -\infty$ and we have

$$|Su(x)| \leqq \sup_{\mathbb{R}^N} u + (1-\beta) \tag{2.13}$$

so that Su is also bounded.

Let $u_1, u_2 : \mathbb{R}^N \to \mathbb{R}$ be bounded and $u_1 = u_2 = 0$ in $\mathcal{T}$. For any fixed $\varepsilon > 0$, $x \in \mathbb{R}^N$ we choose a_1 and a_2 such that

$$\inf_{a \in A} \{\beta u_i(x + hb(x, a))\} \geqq \beta u_i(x + hb(x, a_i)) - \varepsilon, \qquad i = 1, 2, \tag{2.14}$$

and then

$$\begin{aligned} Su_1(x) - Su_2(x) &\leqq \beta[u_1(x + hb(x, a_2)) - u_2(x + hb(x, a_2))] + \varepsilon \\ &\leqq \beta \sup_{\Omega} |u_1 - u_2| + \varepsilon, \end{aligned} \tag{2.15}$$

where the last inequality holds because $u_1 = u_2$ on $\mathcal{T}$.

In the same way we prove the converse inequality and, for $\varepsilon \to 0^+$, we obtain

$$\sup_{\Omega} |Su_1 - Su_2| \leqq \beta \sup_{\Omega} |u_1 - u_2|. \tag{2.16}$$

Since $\beta < 1$ by definition, we conclude that there exists at most one bounded solution of (2.11). We already noted that v_h verifies (HJ_h) for any $x \in \Omega$, and by definition $v_h(x) \equiv 0$ for all $x \in \mathcal{T}$. □

The last result of this section says that the discrete version of the minimum time problem can be completely solved in feedback form once a solution of the discrete Hamilton-Jacobi-Bellman boundary value problem (1.9) (i.e., (2.11)) is known.

COROLLARY 2.4. *Let $u : \mathbb{R}^N \to \mathbb{R}$ be a bounded solution of* (2.11). *Then there exists $F : \mathbb{R}^N \setminus \mathcal{T} \to A$ such that*

$$u(x + hb(x, F(x))) = \inf_{a \in A} u(x + hb(x, a)), \tag{2.17}$$

and any such F has the property that the solution z_j of

$$\begin{cases} z_{j+1} = z_j + hb(z_j, F(z_j)) \\ z_0 = x \end{cases}$$

is an optimal trajectory, i.e., the sequence $F(z_j)$ is an optimal control:

$$v_h(x) = J_x^h(\{F(z_j)\}).$$

Proof. By Theorem 2.3, $u(x) = v_h(x)$. Since this function takes values in a discrete set, the inf in (2.17) is a min, which proves the existence of F satisfying (2.17). The Bellman equation in (2.11) implies

$$u(z_k) = \beta u(z_{k+1}) + 1 - \beta \quad \text{for all } k < n_h(x, F(z_j)),$$

which gives easily, also using the boundary condition $u = 0$ on $\mathcal{T}$,

$$u(x) = 1 - \beta^{n_h(x, F(z_j))},$$

which is the desired equality. □

3. A general convergence theorem. We introduce the following notation:

$$\underline{v}(x) = \liminf_{\substack{h \to 0^+ \\ y \to x}} v_h(y), \qquad \bar{v}(x) = \limsup_{\substack{h \to 0^+ \\ y \to x}} v_h(y) \tag{3.1}$$

and note that these functions are defined everywhere in $\mathbb{R}^N$ since $0 \leqq v_h \leqq 1$. For any bounded function $u : \mathbb{R}^N \to \mathbb{R}$ we define

$$u^{\#}(x) = \liminf_{y \to x} u(y), \qquad u^*(x) = \limsup_{y \to x} u(y) \tag{3.2}$$

which are, respectively, the lower semicontinuous and the upper semicontinuous envelopes of u. These functions will play a crucial role in the following lemma.

LEMMA 3.1. *If* (A1) *holds, then $\bar{v}$ (respectively, $\underline{v}$) is a viscosity subsolution (respectively, supersolution) for*

$$\sup_{a \in A} \{v - \nabla v \cdot b(x, a)\} = 1 \quad \text{in} \quad \Omega. \tag{HJB}$$

Proof. Let $\phi \in C^1(\Omega)$ and $x_0 \in \Omega$ be a strict local maximum for $\bar{v} - \phi$. By Lemma A.3 in Barles and Perthame [5], any sequence x_h of maximum points for $v_h^* - \phi$ in $\overline{B(x_0, r)}$ satisfies

$$\lim_{h \to 0^+} x_h = x_0, \qquad \lim_{h \to 0^+} v_h^*(x_h) = \bar{v}(x_0). \tag{3.3}$$

Then there is $h_1>0$ such that $x_h \in B(x_0, r/2)$ for any $h \leqq h_1$ and, by (A1), there exists $h_2>0$ such that $|hb(x_h, a)|<r/2$ for any $h \leqq h_2$. Then we have

$$v_h^*(x_h)-\phi(x_h) \geqq v_h^*(x_h+hb(x_h, a))-\phi(x_h+hb(x_h, a)), \quad \forall h \leqq \bar{h}, \quad \forall a \in A \tag{3.4}$$

and $\bar{h} \equiv \min\{h_1, h_2\}$. Let $x_h^n \to x_h$ be such that

$$\lim_{n\to+\infty} v_h(x_h^n) = v_h^*(x_h). \tag{3.5}$$

Let us fix a control a and take a subsequence, still denoted x_h^n, such that

$$\lim_{n\to+\infty} v_h(x_h^n+hb(x_h^n, a)) = \gamma. \tag{3.6}$$

By definition (3.2), $\gamma \leqq v_h^*(x_h+hb(x_h, a))$, then passing to the limit over the subsequence in the (DDPP) we obtain

$$v_h^*(x_h)-\beta v_h^*(x_h+hb(x_h, a))-1+\beta \leqq 0, \quad \forall a \in A. \tag{3.7}$$

By (3.4), we have

$$\begin{aligned} 0 &\geqq \sup_{a\in A}\{\beta[v_h^*(x_h)-v_h^*(x_h+hb(x_h, a))]+(1-\beta)v_h^*(x_h)-1+\beta\} \\ &\geqq \sup_{a\in A}\{\beta[\phi(x_h)-\phi(x_h+hb(x_h, a))]+(1-\beta)v_h^*(x_h)-1+\beta\}. \end{aligned} \tag{3.8}$$

Then dividing by h and passing to the limit for $h \to 0^+$ (and remembering that $\beta \equiv e^{-h}$) we prove

$$\sup_{a\in A}\{-\nabla\phi(x_0)\cdot b(x_0, a)+\bar{v}(x_0)-1\} \leqq 0 \tag{3.9}$$

that is, $\bar{v}$ is a viscosity subsolution.

Now let us consider the case when $\underline{v}-\phi$ has a strict minimum at $x_0 \in \Omega$. If x_h is a sequence of strict minimum points for $v_h^\#-\phi$ belonging to $\overline{B(x_0, r)}$, we have

$$\lim_{h\to0^+} x_h = x_0, \qquad \lim_{h\to0^+} v_h^\#(x_h) = \underline{v}(x_0). \tag{3.10}$$

Moreover for any ε and $x \in \Omega$, there exists $a_x \in A$ such that

$$v_h(x)-\beta v_h(x+hb(x, a_x))-1+\beta \geqq -\varepsilon. \tag{3.11}$$

Let $x_h^n \to x_h$ be such that $v_h(x_h^n) \to v_h^\#(x_h)$. By the hypotheses on b the sequence $\{b_h^n\}_n \equiv \{b(x_h^n, a_{x_h^n})\}_n$ is bounded and we can write

$$b_h^n = b(x_0, a_{x_h^n})+E(h, n)$$

where $|E(h, n)| \leqq L|x_h^n-x_0|$. Taking a subsequence (still denoted by x_h^n) and passing to the limit for $n \to +\infty$ we get

$$b(x_0, a_{x_h^n}) \to \bar{b}_h \in Q \equiv \overline{\text{co}}\{b(x_0, a): a \in A\}$$

$$E(h, n) \to \bar{E}(h) \quad \text{and} \quad |\bar{E}(h)| \leqq L|x_h-x_0|.$$

There will also be a subsequence (of the one already considered) such that

$$v_h(x_h^n+hb_h^n) \to \gamma \geqq v_h^\#(x_h+h\bar{b}_h+h\bar{E}(h)).$$

Then, by (3.11), passing to the limit for $n \to +\infty$ and $\varepsilon \to 0$ over the subsequence we get

$$v_h^\#(x_h)-\beta v_h^\#(x_h+h\bar{b}_h+h\bar{E}(h)) \geqq 1-\beta. \tag{3.12}$$

Since x_h is a local minimum point for $v_h^{\#} - \phi$ and $\bar{x}_h \equiv x_h + h\bar{b}_h + h\bar{E}(h) \in B(x_h, r_0)$, for h sufficiently small, we have

$$v_h^{\#}(x_h) - v_h^{\#}(\bar{x}_h) \leqq \phi(x_h) - \phi(\bar{x}_h).$$

Substituting the above inequality in (3.12) we obtain

$$\beta[\phi(x_h) - \phi(\bar{x}_h)] + (1-\beta)v_h^{\#}(x_h) \geqq 1 - \beta. \tag{3.13}$$

By Taylor's expansion we have

$$\phi(x_h) - \phi(\bar{x}_h) = \nabla\phi(x_h) \cdot (h\bar{b}_h + h\bar{E}(h)) + o(h) = \nabla\phi(x_0) \cdot h\bar{b}_h + o(h).$$

Then dividing by h in (3.13), extracting a subsequence such that $\bar{b}_h \to b_0 \in Q$, and passing to the limit as $h \to 0$, we obtain

$$-\nabla\phi(x_0) \cdot b_0 + \underline{v}(x_0) \geqq 1. \tag{3.14}$$

Since $\sup_{b \in \overline{\text{co}}K} p \cdot b = \sup_{b \in K} p \cdot b$ we get

$$\sup_{a \in A} \{-\nabla\phi(x_0) \cdot b(x_0, a) + \underline{v}(x_0)\} \geqq 1$$

that is, $\underline{v}$ is a viscosity super-solution. □

We remark that, by definition, $\underline{v} \leqq \bar{v}$ on Ω. Now we want to prove the converse inequality. We define

$$d(x) \equiv \text{dist}\,(x, \partial\mathcal{T})$$

and

$$X_\delta \equiv \{x\colon \text{dist}\,(x, \partial X) < \delta\}.$$

THEOREM 3.2. *Let v_1(respectively, v_2)$:\mathbb{R}^N \to \mathbb{R}$ be an upper semicontinuous (respectively, lower semicontinuous) bounded viscosity subsolution (respectively, super-solution) of*

$$v + \sup_{a \in A} \{-\nabla v \cdot b(x, a)\} - 1 = 0 \text{ in } \Omega \tag{3.15}$$

such that for some $\delta > 0$

$$|v_i(x)| \leqq \omega(d(x)) \qquad i = 1, 2 \quad \text{for all } x \in \mathcal{T}_\delta, \tag{3.16}$$

where ω tends to 0 as its argument goes to 0. Then $v_1 \leqq v_2$ in $\mathbb{R}^N$.

Proof. Condition (3.16) implies that v_1 and v_2 are both continuous on $\partial\mathcal{T}$ and $v_1(x) = v_2(x) = 0$, for all $x \in \partial\mathcal{T}$. Then the proof of Theorem 2 of [3] applies. The possibility of comparing semicontinuous functions has been remarked in Crandall, Ishii, and Lions [12]. □

THEOREM 3.3. *Assume* (A1) *and v continuous on $\partial\mathcal{T}$. Suppose there exists a continuous $\sigma:\mathbb{R}_+^2 \to \mathbb{R}_+$ such that $\sigma(0, 0) = 0$ and for some $\delta, \bar{h} > 0$*

$$v_h(x) \leqq \sigma(d(x), h), \quad \text{for all } x \in \mathcal{T}_\delta \quad \text{and} \quad h \leqq \bar{h}. \tag{3.17}$$

Then

$$v(x) = \underline{v}(x) = \bar{v}(x) \quad \text{for all } x \in \mathbb{R}^N,$$

and v_h converge uniformly to v on compact subsets of $\mathbb{R}^N$ as $h \to 0^+$.

Proof. Condition (3.17) implies

$$0 \leqq \underline{v}(x) \leqq \bar{v}(x) \leqq \sigma(2d(x), 0) \equiv \omega(d(x)).$$

Then by Lemma 3.1 and Theorem 3.2 we have

$$\bar{v}(x) \leqq \underline{v}(x) \quad \text{for all } x \in \mathbb{R}^N,$$

and therefore v_h converges uniformly on compact sets to the unique continuous viscosity solution w of (1.6).

Since v is assumed continuous only on $\partial\mathcal{T}$, we cannot yet conclude that $w = v$. By Theorem 3.1 in [23] we know that v^* and $v^\#$ are, respectively, a sub- and a supersolution of (3.15). The continuity of v on $\partial\mathcal{T}$ implies (3.16) for v^* and $v^\#$. Then Theorem 3.2 gives

$$v^* \leqq w \leqq v^\#$$

and the proof is complete. □

Remark 3.4. The continuity of v is an assumption of local controllability of system (1.1) around the whole target, while (3.17) states a sort of discrete local controllability (see Remarks 4.2 and 5.3). In the following two sections we will give a sufficient condition on $\mathcal{T}$ and b for both of these assumptions to hold.

Remark 3.5. There are some interesting problems where neither (3.17) holds nor the minimum time function is continuous, because of a lack of controllability on some part of the target. In these cases the following boundary condition is satisfied on $\partial\mathcal{T}$ (in the viscosity sense):

$$\text{either } v \leqq 0 \quad \text{or} \quad v + \sup_{a \in A} \{-\nabla v \cdot b(x, a)\} - 1 \leqq 0,$$

see [5], [6], [23]. Barles and Perthame [5] have proved a uniqueness theorem for semicontinuous solutions of such a boundary value problem which perhaps can be used to prove some convergence of our scheme even when there is no local controllability. For instance, there may be some hope to obtain the uniform convergence of the approximation scheme away from the discontinuities of v. However this seems a rather hard goal to pursue and we limit ourselves here to the locally controllable case.

4. A sufficient condition for discrete controllability. In this section we introduce the following assumptions on $\mathcal{T}$ and b:

$$\text{(A2)} \quad \begin{cases} \text{(i)} \ \mathcal{T} = \{x : g_i(x) \leqq 0 \quad \forall\, i = 1, \cdots, M\} \text{ where } g_i \in C^2(\mathbb{R}^N) \text{ and} \\ \quad |\nabla g_i(x)| > 0 \text{ for all } x \text{ such that } g_i(x) = 0; \\ \text{(ii)} \ \forall x \in \partial\mathcal{T} \quad \exists\, a \in A \text{ such that } g_i(x) = 0 \text{ implies } b(x, a) \cdot \nabla g_i(x) < 0. \end{cases}$$

Hypothesis (A2) means that $\partial\mathcal{T}$ is piecewise C^2 and at each point of the boundary the controller can choose a vector field pointing inward $\mathcal{T}$. A simple special case of (A2) is

$$\text{(A2')} \quad \begin{cases} \text{(i)} \ \mathcal{T} \text{ is the closure of an open bounded set with } C^2 \text{ boundary,} \\ \text{(ii)} \ \inf_{a \in A} b(x, a) \cdot \eta(x) < 0 \text{ for all } x \in \partial\mathcal{T}, \text{ where } \eta(x) \text{ represents} \\ \quad \text{the outward normal to } \mathcal{T} \text{ at } x. \end{cases}$$

We want to show that these assumptions guarantee that condition (3.17) holds. Assumption (A2) is also sufficient for the local controllability of system (1.1) and the Lipschitz continuity of T, as we will see in the following section.

LEMMA 4.1. *Assume* (A1), (A2). *Then there exist some positive constants* δ, $\bar{h}$, C *such that*

$$hN_h(x) \leqq Cd(x) + h, \quad \forall h < \bar{h} \quad \textit{and} \quad \forall\, x \in \mathcal{T}_\delta. \tag{4.1}$$

Proof. We first give the proof under assumption (A2') in order to show how the constants C, $\bar{h}$ and δ can be computed.

Let $x_0 \in \partial\mathcal{T}$ and choose r_0, ξ_0, and $a_0 \in A$ such that d can be redefined as a negative function inside $\mathcal{T}$ to be C^2 in $B_0 \equiv B(x_0, r_0)$ and

$$b(x, a_0) \cdot \eta(x) \leqq -\xi_0 < 0, \quad \forall\, x \in B_0, \tag{4.2}$$

where we have extended $\eta(x)$ out of $\partial\mathcal{T}$ by setting $\eta(x) \equiv \nabla d(x)$. Define

$$C_1 \equiv \sup_{B_0} |D^2 d|, \qquad C_2 \equiv \sup_{B_0 \times A} |b(x, a)| \leqq K(1 + |x_0| + r_0).$$

Since any trajectory of (2.5) living in B_0 satisfies $|x_j - x| \leqq jhC_2$, it is easy to see that

$$x_j \in B_0 \text{ if } x \in B\left(x_0, \frac{r_0}{2}\right) \quad \text{and} \quad jh \leqq \frac{r_0}{3C_2}. \tag{4.3}$$

The solution of (2.5) corresponding to the constant control a_0 satisfies, as long as $x_k \in B_0$ for $k = 0, 1, \cdots, j$,

$$d(x_j) \leqq d(x_{j-1}) + hb(x_{j-1}, a_0) \cdot \nabla d(x_{j-1}) + h^2 C_1 C_2^2 \leqq d(x) - jh\xi_0 + jh^2 C_3,$$

where $C_3 \equiv C_1 C_2^2$ and we have used Taylor's expansion of $d(x)$ and (4.2). If we restrict to

$$h \leqq h_1 \equiv \frac{\xi_0}{2C_3},$$

we get

$$d(x_j) \leqq d(x) - \tfrac{1}{2} jh\xi_0, \quad \text{if } x_k \in B_0, \qquad k = 0, 1, \cdots, j. \tag{4.4}$$

Now if we take

$$j^* = \left[\frac{2\, d(x)}{\xi_0 h}\right] + 1,$$

(4.3) and (4.4) imply $d(x_j^*) \leqq 0$ provided $x \in B(x_0, r_0/2)$, $j^* h \leqq r_0/(3C_2)$ and the last condition holds if

$$d(x) \leqq \delta_1 \equiv \frac{r_0 \xi_0}{12 C_2}, \qquad h \leqq h_2 \equiv \frac{r_0}{6C_2}.$$

Then

$$hN_h(x) \leqq \frac{2d(x)}{\xi_0} + h \quad \forall\, x \in B(x_0, \delta_0), \quad h \leqq h_0, \tag{4.5}$$

where $\delta_0 \equiv \min\{r_0/2, \delta_1\}$, $h_0 \equiv \min\{h_1, h_2\}$.

By the compactness of $\partial\mathcal{T}$ we can cover a neighbourhood of $\mathcal{T}$ by a finite number of balls where (4.5) holds and then get (4.1) with C, δ, $\bar{h}$ given by the minimum of the corresponding constants $(2/\xi_0)$, δ_0, h_0.

We now show how the above arguments work in the general case (A2). We begin multiplying the g_i for suitable positive constants so that in a large compact set containing $\mathcal{T}$ they satisfy

$$g_i(x) \leqq \operatorname{dist}(x, \{g_i(x) \leqq 0\}) \leqq d(x), \qquad i = 1, \cdots, M. \tag{4.6}$$

Given $x_0 \in \partial\mathcal{T}$, let I be the maximal subset of $\{1, \cdots, M\}$ such that $g_i(x_0) = 0$ for all $i \in I$. We first choose r_1 such that

$$B(x_0, r_1) \cap \mathcal{T} = B(x_0, r_1) \cap \{x\colon g_i(x) \leqq 0 \quad \forall\, i \in I\}.$$

Now we take $r_0 < r_1$, ξ_0 and a_0 such that

$$b(x, a_0) \cdot \nabla g_i(x) \leqq -\xi_0 < 0, \quad \forall\, x \in B_0,\ i \in I.$$

We are going to repeat for each g_i, $i \in I$, the calculations made for d. We arrive at

$$g_i(x_j) \leqq g_i(x) - \tfrac{1}{2} jh\xi_0, \quad \text{if } x_k \in B_0, \quad k = 0, \cdots, j, \quad i \in I, \quad h \leqq h_1,$$

where $C_1 \equiv \sup\{|D^2 g_i(x)|\colon x \in B_0 \text{ and } i \in I\}$ and h_1 is modified accordingly. Then for $d(x) \leqq \delta_1$, $h \leqq h_2$ and j^* chosen as above, using (4.6) we get

$$x_{j^*} \in B_0 \quad \text{and} \quad g_i(x_{j^*}) \leqq 0 \quad \text{for } i \in I.$$

This means that $x_{j^*} \in \mathcal{T}$, thus (4.5) holds and we conclude as before. □

Remark 4.2. Assumption (A2) gives also a sort of discrete local controllability. Let us consider the sets

$$\mathcal{R}_h(j) = \{x \in \mathbb{R}^N\colon \exists\{a_i\} \text{ such that } x_j \in \mathcal{T}\}, \quad j \in \mathbb{N}.$$

Obviously

$$\mathcal{R}_h = \bigcup_{j \in \mathbb{N}} \mathcal{R}_h(j).$$

It is easy to see from the proof of Lemma 4.1 that for each h and j, there exists $\delta_j(h) > 0$ such that $\mathcal{T}_{\delta_j(h)} \subseteq \mathcal{R}_h(j)$. This means that (A2) is a sufficient condition for

$$(LC_h) \qquad \mathcal{T} \subseteq \mathring{\mathcal{R}}_h(j) \quad \forall\, j \in \mathbb{N}, \quad \forall\, h < \bar{h}.$$

Note that (LC_h) is a discrete analogue of the local controllability condition for the continuous problem (see Remark 5.3).

5. A necessary and sufficient condition for the Lipschitz continuity of T. In this section we show that (A2) is also a sufficient condition for the controllability of the system (1.1) around $\mathcal{T}$, which gives the continuity of T. Therefore Theorem 3.3 applies and so (A2) is a sufficient condition for the uniform convergence of v_h to v. We also prove that if $\partial\mathcal{T}$ is smooth, i.e., (A2′)(i) holds, then condition (A2′)(ii) is necessary and sufficient for the local Lipschitz continuity of T in $\mathcal{R}$. Finally we compare (A2′) with the classical "positive basis condition" by Petrov for the case $\mathcal{T} = \{0\}$ and discuss the applicability of our approximation results to such a case.

LEMMA 5.1. *Assume* (A1), (A2). *Then there exist* δ', $C > 0$ *such that for all* $x \in \mathcal{T}_{\delta'}$ *there exists a constant control* $\alpha(t) \equiv a$ *such that*

$$t_x(\alpha) \leqq C\, d(x).$$

In particular

$$(5.1) \qquad T(x) \leqq C\, d(x) \quad \forall\, x \in \mathcal{T}_{\delta'}.$$

Proof. We follow the arguments of the proof of Lemma 4.1. Given $x_0 \in \partial\mathcal{T}$ we define I, r_0, ξ_0, a_0, C_1, C_2 as in that proof. For any trajectory of (1.1) in B_0 we have

$$(5.2) \qquad |y_x(s) - x| \leqq sC_2$$

which implies

$$y_x(s) \in B_0 \quad \text{if } x \in B\left(x_0, \frac{r_0}{2}\right) \quad \text{and} \quad s \leqq \frac{r_0}{3C_2}.$$

For such s and the constant control $\alpha(t) \equiv a_0$ we have

$$\begin{aligned} g_i(y_x(s)) &\leqq g_i(x) + \int_0^s \nabla g_i(x) \cdot b(y_x(t), a_0)\, dt + C_1 C_2^2 s^2 \\ &\leqq g_i(x) - \xi_0 s + C_3 s^2, \quad i \in I, \end{aligned}$$

where $C_3 \equiv LC_2 \sup\{|\nabla g_i(x)|: x \in B_0, i \in I\} + C_1 C_2^2$, and in the last inequality we have used (A1) and (5.2). Then

$$g_i(y_x(s)) \leqq g_i(x) - \frac{1}{2}\xi_0 s, \quad \text{if } i \in I, \quad s \leqq s_0 \equiv \min\left\{\frac{r_0}{3C_2}, \frac{\xi_0}{2C_3}\right\}. \tag{5.3}$$

Now $s^* \equiv 2\,d(x)/\xi_0 \leqq s_0$, provided $d(x) \leqq \delta_1 \equiv \xi_0 r_0/(6C_2)$ and $d(x) \leqq \delta_2 \equiv \xi_0^2/(4C_3)$. Then we set $\delta_0 \equiv \min\{r_0/2, \delta_1, \delta_2\}$, and combining (5.3) and (4.6) we obtain

$$y_x(s^*) \in B_0 \text{ and } g_i(y_x(s^*)) \leqq 0 \quad \text{for } i \in I, \quad x \in B(x_0, \delta_0).$$

This means that $y_x(s^*) \in \mathcal{T}$ and gives

$$t_x(\alpha) \leqq \frac{2\,d(x)}{\xi_0} \quad \forall\, x \in B(x_0, \delta_0).$$

A compactness argument completes the proof. □

Combining Theorem 3.3, Lemma 4.1, and Lemma 5.1 we obtain one of the main results of the paper.

THEOREM 5.2. *Assume* (A1), (A2). *Then* $v_h \to v$ *locally uniformly in* $\mathbb{R}^N$ *and* $hN_h \to T$ *locally uniformly in* $\mathcal{R}$.

Proof. Under the above assumptions,

$$v_h(x) = 1 - e^{-hN_h(x)} \leqq 1 - e^{-(Cd(x)+h)} \leqq C\,d(x) + h.$$

Then Theorem 3.3 applies for $\sigma(d(x), h) \equiv C\,d(x) + h$. □

Remark 5.3. Lemma 5.1 immediately provides the following form of small time local controllability: if $\mathcal{R}(t)$ is the set of points from which the system can reach $\mathcal{T}$ in time smaller than t, i.e.,

$$\mathcal{R}(t) := \{x: T(x) < t\},$$

then we have

$$\mathcal{T} \subseteq \mathring{\mathcal{R}}(t), \quad \text{for all } t > 0.$$

This result has already been proved in a more general context by Bacciotti and Stefani [2]. Results of this type have been studied extensively in the case $\mathcal{T} = \{0\}$, see e.g., Petrov [28], Lee and Markus [25], Hermes and La Salle [21], Bacciotti [1], Sussmann [31].

THEOREM 5.4. *Under the assumptions* (A1), (A2) *the minimum time function* T *is locally Lipschitz continuous in* $\mathcal{R}$.

Proof. Let us take x, z in a compact subset of $\mathcal{R}$ and suppose $T(x) \leqq T(z) < \tilde{T}$. We fix $\varepsilon > 0$ and a control α such that

$$T(x) + \varepsilon \leqq \tilde{T} \quad \text{and} \quad y_x(T(x)+\varepsilon) \equiv x_1 \in \mathcal{T}.$$

Let $y_z(t)$ be the trajectory starting at z under the same control α and $z_1 \equiv y_z(T(x)+\varepsilon)$. By Gronwall's lemma we have

$$|x_1 - z_1| \leqq e^{L\tilde{T}}|x - z|.$$

In order to apply Lemma 5.1 we choose

$$|x - z| \leqq \delta' e^{-L\tilde{T}},$$

and find

$$T(z) - T(x) \leqq \varepsilon + T(z_1) \leqq \varepsilon + C\,e^{L\tilde{T}}|x - z|,$$

and the conclusion follows from the arbitrariness of ε. □

The next theorem will show, in particular, that (A2′)(ii) is a necessary condition for the Lipschitz continuity of T in a neighbourhood of $\mathcal{T}$ when $\mathcal{T}$ is smooth.

THEOREM 5.5. *Assume* (A1) *and* (A2')(i). *Suppose there exist positive constants* C, $\bar{r}$, *and* $\frac{1}{2} < \alpha \leqq 1$ *such that*

$$T(x) \leqq C\, d^{\alpha}(x), \quad \textit{for all } x \in B(\bar{x}, \bar{r}). \tag{5.4}$$

Then

$$\inf_{a \in A} b(x, a) \cdot \eta(x) < 0, \quad \textit{for all } x \in \partial \mathcal{T} \cap B\left(\bar{x}, \frac{\bar{r}}{2}\right), \tag{5.5}$$

where η *is the exterior normal to* $\mathcal{T}$.

Proof. We assume by contradiction that the inequality in (5.5) is violated for some $x_0 \in \partial \mathcal{T} \cap B(\bar{x}, \bar{r}/2)$, so that by (A1)

$$\inf_{a \in A} b(x, a) \cdot \eta(x) \geqq -L|x - x_0|, \quad \text{for all } x \in \mathcal{T}_{\delta}, \tag{5.6}$$

where δ is such that $d \in C^2(\mathcal{T}_{\delta} \setminus \mathcal{T})$, and we have defined $\eta(x) \equiv \nabla d(x)$. We fix $\varepsilon > 0$, $x_n \to x_0$ and controls such that

$$y_{x_n}(T(x_n) + \varepsilon) \equiv y_n \in \mathcal{T}. \tag{5.7}$$

Observe that $T(x_n) + \varepsilon \leqq \tilde{T}$, for all n, and (A1) imply

$$|y_{x_n}(s)| \leqq (|x_0| + \delta + 1)\, e^{K\tilde{T}} \equiv Y \quad \text{for } s \leqq \tilde{T},$$

and then

$$|y_n - x_n| \leqq (T(x_n) + \varepsilon) K (1 + Y). \tag{5.8}$$

Using a Taylor expansion of $d(x)$, (5.6), (5.7), and (5.8) we obtain

$$\begin{aligned} d(x_n) &\leqq d(y_n) - \int_0^{T(x_n)+\varepsilon} \nabla d(y_n) \cdot b(y_n, \alpha(t))\, dt + O((T(x_n) + \varepsilon)^2) \\ &\leqq L|y_n - x_0|(T(x_n) + \varepsilon) + O((T(x_n) + \varepsilon)^2) \\ &\leqq L|x_n - x_0|(T(x_n) + \varepsilon) + O((T(x_n) + \varepsilon)^2). \end{aligned}$$

Now we make the more precise choice $x_n = x_0 + (1/n)\eta(x_0)$, to have $|x_n - x_0| \leqq 2\, d(x_n)$ for n sufficiently small, and deduce from the previous formula and (5.4)

$$d(x_n) \leqq 2LC\, d^{1+\alpha}(x_n) + O(d^{2\alpha}(x_n)),$$

which gives a contradiction if $\alpha > \frac{1}{2}$. □

Remark 5.6. Theorem 5.5 is of local nature whereas Theorem 5.4 is global, but it is easy to give a local version of it making minor changes to its proof and the proof of Lemma 5.1. More precisely, if $\bar{x} \in \partial \mathcal{T}$, $\partial \mathcal{T} \cap B(\bar{x}, \bar{r})$ is a C^2 manifold with $\mathcal{T}$ lying on one side of it and $\inf_{a \in A} b(x, a) \cdot \eta(x) < 0$ for all $x \in \partial \mathcal{T} \cap B(\bar{x}, \bar{r})$, then $T(x)$ is Lipschitz continuous in $\mathcal{T}_{\delta} \cap B(\bar{x}, (\bar{r}/2))$ for some $\delta > 0$. A similar result can be given for a piecewise C^2 portion of $\partial \mathcal{T}$.

Remark 5.7. (*The relation between* (A2) *and Petrov's condition*). Condition (A2) (ii) is the counterpart for piecewise smooth targets of the classical "positive basis condition" due to Petrov, which is necessary and sufficient for the Lipschitz continuity of T in the case $\mathcal{T} = \{0\}$ [29]. One of the equivalent formulations of this condition is the following [28]:

(PC) for every unit vector γ there exist $a \in A$ such that $b(0, a) \cdot \gamma < 0$.

The explicit relation between (PC) and (A2) is summarized in the following statements whose proof is straightforward:

(i) if (PC) holds, then there exists $\varepsilon > 0$ such that every $\mathcal{T} \subseteq B(0, \varepsilon)$ and satisfying (A2)(i), must satisfy (A2)(ii);

(ii) if there exist $\xi, \bar{\varepsilon} > 0$ such that for all $\varepsilon < \bar{\varepsilon}$

$$\inf_{a \in A} b(x, \alpha) \cdot \eta(x) < -\xi \quad \text{for all } x \in \partial B(0, \varepsilon),$$

then (PC) holds. If (A2′, i) holds, then Theorem 5.4 is a special case of a result by Friedman [32] on differential games of pursuit and evasion. The relation between Friedman's assumption, i.e., (A2′, ii), and the positive basis condition (PC), was remarked by Petrov in [33].

Remark 5.8. (*Approximation of T in the case* $\mathcal{T} = \{0\}$). In order to apply the convergence results to the case $\mathcal{T} = \{0\}$ we need a further approximation step. We consider a family of approximating targets $\mathcal{T}_\varepsilon \equiv B(0, \varepsilon)$ with unchanged dynamics and denote by $\mathcal{R}_\varepsilon$ and T_ε the corresponding controllable set and minimum time function. If Petrov's condition (PC) holds then by Petrov's Theorem [29] there exist $\bar{\varepsilon} > 0$, and a positive constant C_1 such that $B(0, \bar{\varepsilon}) \subset \mathcal{R}$ and

$$T(x) \leqq C_1 |x| \quad \forall\, x \in B(0, \bar{\varepsilon}).$$

Therefore it is easy to show that, for any $\varepsilon \leqq \bar{\varepsilon}$,

$$\mathcal{R}_\varepsilon = \mathcal{R}$$

and

$$0 \leqq T(x) - T_\varepsilon(x) \leqq C_1 \varepsilon \quad \forall\, x \in \mathcal{R}.$$

Moreover, by Remark 5.7(i), we can choose $\bar{\varepsilon}$ such that (A2) is satisfied for any approximating problem with $\varepsilon < \bar{\varepsilon}$. Then the convergence theorem applies to each of these problems and the discretization method can be used to obtain an approximate solution of the minimum time problem in the case $\mathcal{T} = \{0\}$ as well.

6. Appendix. A brief comparison with other approximation schemes. Several numerical methods for the minimum time problem have been proposed, mostly for linear systems, point-shaped targets, and under global controllability assumptions. A first group of papers [27], [14], [19] exploit the explicit representation of the trajectories of the controlled system available in the linear case, which provides a bang-bang optimal control defined componentwise by

$$\alpha(t) \equiv \operatorname{sign}(Y(t) \cdot \eta^*) \quad \text{for all } t < t^*, \tag{6.1}$$

where Y is a matrix which can be computed from the linear vector field, t^* is the minimum time, and η^* is some constant vector. Then the minimum time problem reduces to finding an appropriate η^* to substitute in (6.1). Neustadt [27] showed that, for a fixed point-shaped target and under "normality conditions" (see [21]), η^* is the point where a suitable regular function attains its maximum so that a gradient method can be used to compute it. Eaton [14] extended this procedure to the case of a moving target. He used some geometric properties of the reachable sets for normal linear systems to give an iterative procedure which computes $\{t_k\}$ and $\{\eta_k\}$ converging, respectively, to t^* and η^*. Finally Fujisawa and Yasuda [19] proved that, without normality conditions, a modification of Eaton's procedure gives exponential convergence of t_k toward t^* for fixed point-shaped targets.

We want to emphasize several important differences between these methods and ours.

(1) All these methods do not apply in the nonlinear case where a formula like (6.1) for an optimal control is missing. On the other hand they can handle the nonautonomous case.

(2) The above procedures compute an optimal strategy steering a given initial point x_0 to the final state. For a different x_0 the algorithm must be restarted. On the contrary, dynamic programming gives information for all initial points in a domain.

(3) The optimal control computed by applying these methods is in open loop form, whereas the dynamic programming approach provides a feedback control (see Corollary 2.4).

(4) Using these techniques it does not seem possible to identify whether an initial point is controllable to the target in finite time, i.e., to know if $x_0 \in \mathcal{R}$ or not. On the contrary our approach is global and gives at least an approximate knowledge of $\mathcal{R}$.

(5) The above procedures do not seem adequate to treat general targets instead of a single terminal point.

A completely different method can be found in the book by Canon, Cullum, and Polak [7] (see also the references therein), again for linear systems satisfying a controllability assumption. It considers a discrete version of the system and derives a sequence of linear programming problems of growing dimension, each one solved by the simplex algorithm. The discrete version of the system corresponds to the Euler scheme (2.5) with the choice $h = 1$, but the convergence to the continuous-time system as $h \to 0$ is not studied. It is worthwhile to note that a slight modification of the method can also be used to determine whether a point can be steered to the final state in a given number of steps. The remarks (1), (2), (3), and (5) still hold. In particular, the solution of the minimum time problem in a given set instead of that for a single initial point may lead to a very large amount of computations.

Finally, a method for treating the problem in the nonlinear case can be found in the book by Falb and de Jong [16] (see also the references therein). Under strong regularity assumptions on the data, the Pontryagin minimum principle is used to obtain a two point boundary value problem solvable by means of the Newton-Kantorovich method in Banach spaces. The convergence of this method is sensitive to the choice of the initial guess. Remarks (2), (3), (4), and (5) remain valid for this method.

Acknowledgments. We wish to thank our colleagues of CEREMADE for their kind hospitality.

REFERENCES

[1] A. BACCIOTTI, *Fondamenti geometrici della teoria della controllabilita'*, Quaderno U.M.I. n.31, Pitagora edition, Bologna, 1986.

[2] A. BACCIOTTI AND G. STEFANI, *Self-accessibility of a set with respect to a multivalued field*, J. Optim. Theory Appl., 31 (1980), pp. 535-552.

[3] M. BARDI, *A boundary value problem for the minimum time function*, SIAM J. Control Optim., 27 (1989), pp. 776-785.

[4] M. BARDI AND M. FALCONE, *Discrete approximation of the minimum time function for systems with regular optimal trajectories*, preprint, 1989.

[5] G. BARLES AND B. PERTHAME, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Math. Modelling and Num. Anal., 21 (1987), pp. 557-579.

[6] ———, *Exit time problems in optimal control and vanishing viscosity method*, SIAM J. Control Optim., 26 (1988), pp. 1133-1148.

[7] M. D. CANON, C. D. CULLUM, AND E. POLAK, *Theory of Optimal Control and Mathematical Programming*, McGraw Hill, New York, 1970.

[8] I. CAPUZZO DOLCETTA, *On a discrete approximation of the Hamilton-Jacobi equation of dynamic programming*, Appl. Math. Optim., 10 (1983), pp. 367-377.

[9] I. CAPUZZO DOLCETTA AND M. FALCONE, *Viscosity solutions and discrete dynamic programming*, Ann. Inst. H. Poincaré, Anal. Non Lin., 6 (Supplement) (1989), pp. 161-183.

[10] I. CAPUZZO DOLCETTA AND H. ISHII, *Approximate solutions of the Bellman equation of deterministic control theory*, Appl. Math. Optim., 11 (1984), pp. 161-181.

[11] R. CONTI, *Processi di controllo lineari in* $\mathbb{R}^N$, Quaderno U.M.I. n. 30, Pitagora edition, Bologna, 1985.

[12] M. G. CRANDALL, H. ISHII, AND P. L. LIONS, *Uniqueness of viscosity solutions of Hamilton-Jacobi equations revisited*, J. Math. Soc. Japan, 39 (1987), pp. 581-596.

[13] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[14] J. H. EATON, *An iterative solution to time-optimal control*, J. Math. Anal. Appl., 5 (1962), pp. 329-344.

[15] L. C. EVANS AND M. R. JAMES, *The Hamilton-Jacobi-Bellman equation for time-optimal control*, preprint, 1988.

[16] P. L. FALB AND J. L. DE JONG, *Some successive approximation methods in control and oscillation theory*, Academic Press, New York, 1969.

[17] M. FALCONE, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 1-13.

[18] W. H. FLEMING AND R. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, 1975.

[19] T. FUJISAWA AND Y. YASUDA, *An iterative procedure for solving the time-optimal regulator problem*, SIAM J. Control, 5 (1967), pp. 501-512.

[20] H. HERMES, *Feedback synthesis and positive, local solutions to Hamilton-Jacobi-Bellman equations*, Proc. MTNS 87, C. I. Byrnes, C. F. Martin, and R. E. Sacks, eds., North Holland, Amsterdam, 1988.

[21] H. HERMES AND J. P. LA SALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[22] H. ISHII, *Perron's method for Hamilton-Jacobi equations*, Duke Math. J., 55 (1987), pp. 369-384.

[23] ———, *A boundary value problem of the Dirichlet type for Hamilton-Jacobi equations*, Ann. Scuola Norm. Sup. Pisa, to appear.

[24] S. N. KRUZKOV, *Generalized solutions of the Hamilton-Jacobi equations of eikonal type* I, Math. USSR Sbornik, 27 (1975), pp. 406-445.

[25] E. B. LEE AND L. MARKUS, *Foundations of optimal control*, John Wiley, New York, 1967.

[26] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, London, 1982.

[27] L. W. NEUSTADT, *Synthesizing time optimal control systems*, J. Math. Anal. Appl., 1 (1960), pp. 484-493.

[28] N. N. PETROV, *Controllability of autonomous systems*, Differential Equations, 4 (1968), pp. 606-617.

[29] ———, *On the Bellman function for the time-optimal process problem*, J. Appl. Math. Mech., 34 (1970), pp. 785-791.

[30] P. E. SOUGANIDIS, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, J. Differential Equations, 59 (1985), pp. 1-43.

[31] H. J. SUSSMAN, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158-194.

[32] A. FRIEDMAN, *Existence of value and of saddle points for differential games of pursuit and evasion*, J. Differential Equations, 7 (1970), pp. 92-110.

[33] N. N. PETROV, *On the continuity of the Bellman function with respect to a parameter*, Vestnik Leningrad Univ. Math., 7 (1979), pp. 169-176.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 966-979, July 1990

011

A GENERAL STOCHASTIC MAXIMUM PRINCIPLE FOR OPTIMAL CONTROL PROBLEMS*

SHIGE PENG†

Abstract. The maximum principle for nonlinear stochastic optimal control problems in the general case is proved. The control domain need not be convex, and the diffusion coefficient can contain a control variable.

Key words. stochastic optimal control, maximum principle, variational inequality

AMS(MOS) subject classifications. 93E, 60H

1. Introduction. In this paper we study the following type of stochastic optimal control problem. Minimize a cost function

$$J(v(\cdot)) = E\int_0^T l(x(t), v(t))\,dt + Eh(T)$$

subject to

$$dx(t) = g(x(t), v(t))\,dt + \sigma(x(t), v(t))\,dB(t),$$
$$x(0) = x_0.$$

In the above, $v(\cdot)$ is the control variable valued in a subset of R^k, $x(\cdot)$ is the state variable, $B(\cdot)$ is a standard Wiener process, and l, h, g, σ are given maps. Our object is to obtain a necessary condition, called the maximum principle, for optimal control. There are many works concerning this subject (see [1]-[4], [7], [8], [10]). A difficulty is treating the case where the diffusion coefficient σ contains the control variable v. Bensoussan [1], [4] studied such a case. The maximum principle he obtains is of local condition (see (26) of this paper), and his method depends heavily on the control being convex. In our problem, since the control domain is not necessarily convex, we must obtain the maximum principle in its global form. A classical way of treating such a problem is to use the "spike variation method" [12]. More precisely, if $u(\cdot)$ is an optimal control and v is arbitrary then we can define an admissible control as follows:

$$u^\varepsilon(t) = \begin{cases} v & \text{if } s \leqq t \leqq s+\varepsilon, \\ u(t) & \text{otherwise,} \end{cases}$$

with a sufficiently small $\varepsilon > 0$. Then, we derive the variational equation from the state equation, and the variational inequality from the inequality

$$J(u^\varepsilon(\cdot)) - J(u(\cdot)) \geqq 0.$$

But in this situation σ contains v, and thus, from the spike variation of control $u(\cdot)$, we can only obtain the following estimation:

$$(*) \qquad E|y^\varepsilon(t) - y(t)|^2 = O(\varepsilon),$$

where $y^\varepsilon(\cdot)$ and $y(\cdot)$ are the trajectories of the state equation corresponding to $u^\varepsilon(\cdot)$ and $u(\cdot)$, respectively. We should note that in [1] and [3], we have the estimation

$$E|x^\varepsilon(t) - y(t)|^2 = O(\varepsilon^2),$$

* Received by the editors February 16, 1989; accepted for publication (in revised form) September 5, 1989.

† Institute of Mathematics, Fudan University, Shanhai, China and Institute of Mathematics, Shandong University, Jinan, China. This work was partially supported by the Chinese National Natural Science Foundation.

where $x^\varepsilon(\cdot)$ is the trajectory of the state equation corresponding to the control $u^\varepsilon(\cdot)=u(\cdot)+\varepsilon v(\cdot)$. Due to the estimation (*), the classical way of deriving the variational equation does not work. We introduce a new approach to overcome this difficulty. The main idea is to consider the second-order terms (with respect to the state) in the expansion of $J(u^\varepsilon(\cdot))-J(u(\cdot))$. Although the sum of these second-order terms are quadratic with respect to the state variable, we can regard it as a linear functional on the product of the state space. Then a so-called second-order variational equation and second-order variational inequality are introduced. Based on this, we obtain the corresponding (second-order) adjoint processes and adjoint equations that lead to the maximum principle. It turns out that the maximum principle we derived is novel, and it contains the earlier work as a special case. The paper is organized as follows. In § 2, we give the statement of the problem and our main assumptions. In § 3, we study the second-order expansion of the perturbed state variable $y^\varepsilon(\cdot)$, and the perturbed cost function $J(u^\varepsilon(\cdot))$. We also treat the estimations of these terms. In § 4, we obtain the first- and second-order adjoint processes. Consequently, the second-order variational inequality is given in this section. Our main result, the maximum principle, is given in § 5. The first- and second-order adjoint equations are also derived in this section. In the last section, we show how to obtain the maximum principle in the case when an endpoint constraint is imposed.

2. Statement of the problem. Let $(\Omega, \mathscr{F}, P)$ be a probability space with filtration $\mathscr{F}^t$. Let $B(\cdot)$ be an R^n-valued standard Wiener process. We assume that

$$\mathscr{F}^t=\sigma\{B(s);\, 0\leqq s\leqq t\}.$$

Consider the following stochastic control system:

$$\begin{aligned} &dx(t)=g(x(t), v(t))\,dt+\sigma(x(t), v(t))\,dB(t),\\ &x(0)=x_0, \end{aligned} \tag{1}$$

where

$$\begin{gathered} g(x, v): R^n\times R^k\to R^n,\\ \sigma(x, v): R^n\times R^k\to \mathscr{L}(R^d, R^n),\\ \sigma=(\sigma^1, \sigma^2, \cdots, \sigma^d). \end{gathered}$$

An admissible control $v(\cdot)$ is an $\mathscr{F}^t$-adapted process with values in U such that

$$\sup_{0\leqq t\leqq T} E|v(t)|^m<\infty \quad \forall m=1, 2, \cdots,$$

where U is a nonempty subset of R^k (control domain). We denote the set of all admissible controls by U_{ad}. Our optimal control problem is to minimize the following cost functional over U_{ad}:

$$\begin{aligned} &J(v(\cdot))=E\int_0^T l(x(t), v(t))\,dt+Eh(x(T)),\\ &\inf\{J(v(\cdot));\, v(\cdot)\in U_{\text{ad}}\}, \end{aligned} \tag{2}$$

where

$$l(x, v): R^n\times R^k\to R, \qquad h(x): R^n\to R.$$

Our assumption is:

(3) g, σ, l, h are twice continuously differentiable with respect to x. They and all their derivatives $g_x, g_{xx}, \sigma_x, \sigma_{xx}, l_x, l_{xx}, h_x, h_{xx}$, are continuous in (x, v).

$g_x, g_{xx}, \sigma_x, \sigma_{xx}, l_{xx}, h_{xx}$ are bounded, and g, σ, l_x, h_x are bounded by $C(1+|x|+|v|)$.

3. Second-order expansion. The purpose of this section is to derive a kind of variational equation and variational inequality. Due to the appearance of the control variable in $\sigma(\cdot,\cdot)$ and the control domain U not necessarily convex, the usual first-order expansion approach does not work. Hence, we introduce a second-order expansion method. Let $(y(\cdot), u(\cdot))$ be an optimal solution of the problem. It is classical to construct a perturbated admissible control in the following way (spike variation):

$$u^\varepsilon(t) = \begin{cases} v & \text{if } \tau \leqq t \leqq \tau+\varepsilon, \\ u(t) & \text{otherwise.} \end{cases}$$

Where $0 \leqq \tau < T$ is fixed, $\varepsilon > 0$ is sufficiently small, and v is an arbitrary $\mathscr{F}^\tau$-measurable random variable with values in U, such that

$$\sup_{\omega\in\Omega} |v(\omega)| < \infty.$$

Let $y^\varepsilon(\cdot)$ be the trajectory of the control system (1) corresponding to the control $u^\varepsilon(\cdot)$. We would like to derive the variational inequality from the fact that

$$J(u^\varepsilon(\cdot)) - J(u(\cdot)) \geqq 0.$$

To this end, we need the following estimation.

LEMMA 1. *We suppose* (3). *Then*

$$\varepsilon^{-2} \sup_{0\leqq t\leqq T} E|y^\varepsilon(t) - y(t) - y_1(t) - y_2(t)|^2 \leqq C \tag{4}$$

where $y_1(\cdot)$, $y_2(\cdot)$ *are solutions of*

$$\begin{aligned} y_1(t) = {} & \int_0^t [g_x(y(s), u(s))y_1(s) + (g(y(s), u^\varepsilon(s)) - g(y(s), u(s)))]\, ds \\ & + \int_0^t [\sigma_x(y(s), u(s))y_1(s) + (\sigma(y(s), u^\varepsilon(s)) - \sigma(y(s), u(s)))]\, dB(s), \end{aligned} \tag{5}$$

$$\begin{aligned} y_2(t) = {} & \int_0^t \left[g_x(y(s), u(s))y_2(s) + \frac{1}{2} g_{xx}(y(s), u(s))y_1(s)y_1(s) \right] ds \\ & + \int_0^t \left[\sigma_x(y(s), u(s))y_2(s) + \frac{1}{2} \sigma_{xx}(y(s), u(s))y_1(s)y_1(s) \right] dB(s) \\ & + \int_0^t (g_x(y(s), u^\varepsilon(s)) - g_x(y(s), u(s)))y_1(s)\, ds \\ & + \int_0^t (\sigma_x(y(s), u^\varepsilon(s)) - \sigma_x(y(s), u(s)))y_1(s)\, dB(s), \end{aligned} \tag{6}$$

where

$$f_{xx}yy = \sum_{i,j=1}^n f_{x^i x^j} y^i y^j \quad \text{for } f = g, \sigma, l, h.$$

Remark. Equation (5) is called the first-order variational equation. It is the variational equation in the usual sense. We must introduce what we call "the second-order variational equation" (6), because with the solution of (5), we can only obtain

the following estimation:

$$\varepsilon^{-1} \sup_{0 \leqq t \leqq T} E|y^{\varepsilon}(t) - y(t) - y_1(t)|^2 \leqq C.$$

It is not enough to derive the variational inequality.

Proof. From the construction of $u^{\varepsilon}(\cdot)$, it is easy to verify by Gronwall's inequality and the moment inequality (see Ikeda and Watanabe [9]) that

$$\sup_{0 \leqq t \leqq T} E|y_1(t)|^2 \leqq C\varepsilon, \tag{7}$$

$$\sup_{0 \leqq t \leqq T} E|y_2(t)|^2 \leqq C\varepsilon^2, \tag{8}$$

$$\begin{cases} \sup\limits_{0 \leqq t \leqq T} E|y_1(t)|^4 \leqq C\varepsilon^2, \\ \sup\limits_{0 \leqq t \leqq T} E|y_2(t)|^4 \leqq C\varepsilon^4, \\ \sup\limits_{0 \leqq t \leqq T} E|y_1(t)|^8 \leqq C\varepsilon^4. \end{cases} \tag{9}$$

Set

$$y_3 = y_1 + y_2.$$

We have

$$\begin{aligned}
&\int_0^t g(y+y_3, u^{\varepsilon})\, ds + \int_0^t \sigma(y+y_3, u^{\varepsilon})\, dB(s) \\
&\quad = \int_0^t [g(y, u^{\varepsilon}) + g_x(y, u^{\varepsilon})y_3 + \int_0^1 \int_0^1 \lambda g_{xx}(y+\lambda\mu y_3, u^{\varepsilon})\, d\lambda\, d\mu\, y_3 y_3]\, ds \\
&\qquad + \int_0^t \left[\sigma(y, u^{\varepsilon}) + \sigma_x(y, u^{\varepsilon})y_3 + \int_0^1 \int_0^1 \lambda\sigma_{xx}(y+\lambda\mu y_3, u^{\varepsilon})\, d\lambda\, d\mu\, y_3 y_3 \right] dB(s) \\
&\quad = \int_0^t g(y, u)\, ds + \int_0^t \sigma(y, u)\, dB(s) + \int_0^t g_x(y, u)y_3\, ds + \int_0^t \sigma_x(y, u)y_3\, dB(s) \\
&\qquad + \int_0^t \Big(g(y(s), u^{\varepsilon}(s)) - g(y(s), u(s)))\, ds \\
&\qquad + \int_0^t (\sigma(y(s), u^{\varepsilon}(s)) - \sigma(y(s), u(s)))\Big)\, dB(s) \\
&\qquad + \int_0^t \frac{1}{2} g_{xx}(y, u)y_3(s)y_3(s)\, ds + \int_0^t \frac{1}{2} \sigma_{xx}(y, u)y_3(s)y_3(s)\, dB(s) \\
&\qquad + \int_0^t (g_x(y, u^{\varepsilon}) - g_x(y, v))y_3(s)\, ds \\
&\qquad + \int_0^t (\sigma_x(y, u^{\varepsilon}) - \sigma_x(y, u))y_3(s)\, dB(s) \\
&\qquad + \int_0^t \int_0^1 \int_0^1 \lambda [g_{xx}(y+\lambda\mu y_3, u^{\varepsilon}) - g_{xx}(y, u)]\, d\lambda\, d\mu\, y_3 y_3\, ds \\
&\qquad + \int_0^t \int_0^1 \int_0^1 \lambda [\sigma_{xx}(y+\lambda\mu y_3, u^{\varepsilon}) - \sigma_{xx}(y, u)]\, d\lambda\, d\mu\, y_3 y_3\, dB(s) \\
&\quad = y(t) + y_3(t) - x_0 + \int_0^t G^{\varepsilon}(s)\, ds + \int_0^t \Lambda^{\varepsilon}(s)\, dB(s),
\end{aligned}$$

where (using (5) and (6))

$$G^\varepsilon(s)=\tfrac{1}{2}g_{xx}(y(s),u(s))(y_2(s)y_2(s)+2y_1(s)y_2(s))$$
$$+(g_x(y(s),u^\varepsilon(s))-g_x(y(s),u(s)))y_2(s)$$
$$+\int_0^1\int_0^1\lambda[g_{xx}(y+\lambda\mu y_3,u^\varepsilon)-g_{xx}(y,v)]\,d\lambda\,d\mu\,y_3(s)y_3(s)$$
$$\Lambda^\varepsilon(s)=\tfrac{1}{2}\sigma_{xx}(y(s),u(s))(y_2(s)y_2(s)+2y_1(s)y_2(s))$$
$$+(\sigma_x(y(s),u^\varepsilon(s))-\sigma_x(y(s),u(s)))y_2(s)$$
$$+\int_0^1\int_0^1\lambda[\sigma_{xx}(y+\lambda\mu y_3,u^\varepsilon)-\sigma_{xx}(y,v)]\,d\lambda\,d\mu\,y_3(s)y_3(s).$$

From (7), (8), and (9) we can check that

$$\sup_{0\leqq t\leqq T}E\left(\left|\int_0^t G^\varepsilon(s)\,ds\right|^2+\left|\int_0^t\Lambda^\varepsilon(s)\,dB(s)\right|^2\right)=o(\varepsilon^2). \tag{10}$$

Thus we have

$$y(t)+y_3(t)=x_0+\int_0^t g(y(s)+y_3(s),u^\varepsilon(s))\,ds$$
$$+\int_0^t\sigma(y(s)+y_3(s),u^\varepsilon(s))\,dB(s)-\int_0^t G^\varepsilon(s)\,ds-\int_0^t\Lambda^\varepsilon(s)\,dB(s).$$

Since

$$y^\varepsilon(t)=x_0+\int_0^t g(y^\varepsilon(s),u^\varepsilon(s))\,ds+\int_0^t\sigma(y^\varepsilon(s),u^\varepsilon(s))\,dB(s),$$

we can derive

$$(y^\varepsilon-y-y_3)(t)=\int_0^t A^\varepsilon(s)(y^\varepsilon-y_3-y)(s)\,ds+\int_0^t D^\varepsilon(s)(y_3^\varepsilon-y-y)(s)\,dB(s)$$
$$+\int_0^t G^\varepsilon(s)\,ds+\int_0^t\Lambda^\varepsilon(s)\,dB(s)$$

with

$$|A^\varepsilon(s,\omega)|+|D^\varepsilon(s,\omega)|\leqq C\quad\forall s,\forall\omega.$$

From this relation and (10), we can use Ito's formula and Gronwall's inequality to obtain the estimation (4). The proof is completed. □

Since $u(\cdot)$ is an optimal control, from Lemma 1 we can easily derive Lemma 2.

LEMMA 2. *Under the assumption of Lemma* 1, *we have*

$$E\int_0^T\left[l_x(y(s),u(s))(y_1(s)+y_2(s))+\frac{1}{2}l_{xx}(y(s),u(s))y_1(s)y_1(s)\right]ds$$
$$+E\int_0^T(l(y(s),u^\varepsilon(s))-l(y(s),u(s)))\,ds$$
$$+E(h_x(y(T))(y_1(T)+y_2(T)))+\tfrac{1}{2}Eh_{xx}(y(T))y_1(T)y_1(T) \tag{11}$$
$$\geqq o(\varepsilon).$$

Remark. In the case where σ does not contain the control variable v, the relation (11) can be reduced to

$$E\int_0^T [l_x(y(s), u(s))y_1(s)\,ds + E(h_x(y(T))y_1(T) + E\int_0^T (l(y(s), u^\varepsilon(s)) - l(y(s), u(s)))\,ds \geqq o(\varepsilon).$$

Thus we need only the first-order variational equation (5).

Proof. Since $(y(\cdot), u(\cdot))$ is optimal, we have

$$E\int_0^T l(y^\varepsilon(t), u^\varepsilon(t))\,dt + Eh(y^\varepsilon(T)) - E\int_0^T l(y(t), u(t))\,dt + Eh(y(T)) \geqq 0.$$

Thus from Lemma 1,

$$\begin{aligned}
0 &\leqq E\int_0^T (l(y+y_1+y_2, u^\varepsilon) - l(y, u))\,dt + E[h(y+y_1+y_2)(T) - h(y(T))] + o(\varepsilon) \\
&= E\int_0^T (l(y+y_1+y_2, u) - l(y, u))\,dt + E[h(y+y_1+y_2)(T) - h(y(T))] \\
&\quad + E\int_0^T (l(y+y_1+y_2, u^\varepsilon) - l(y+y_1+y_2, u))\,dt + o(\varepsilon) \\
&= E\int_0^T \Big[l_x(y(s), u(s))(y_1(s)+y_2(s)) \\
&\quad + \frac{1}{2} l_{xx}(y(s), u(s))(y_1(s)+y_2(s))(y_1(s)+y_2(s)) \Big]\,ds \\
&\quad + E\int_0^T (l(y(s)\cdot u^\varepsilon(s)) - l(y(s), u(s)))\,ds \\
&\quad + E\int_0^T (l_x(y(s)\cdot u^\varepsilon(s)) - l_x(y(s), u(s)))(y_1(s)+y_2(s))\,ds \\
&\quad + \frac{1}{2} E\int_0^T (l_{xx}(y(s), u^\varepsilon(s)) - l_{xx}(y(s), u(s))y_1(s)y_1(s))\,ds \\
&\quad + E(h_x(y(T))(y_1(T)+y_2(T))) \\
&\quad + \frac{1}{2} Eh_{xx}(y(T))y_1(T)y_1(T) + o(\varepsilon).
\end{aligned}$$

Then, (11) follows from (7) and (8). □

4. Adjoint processes and variational inequality. In this section, we introduce the first- and second-order adjoint processes for (5), (6), and (11). With these processes, we can easily derive the variational inequality from (11). The linear terms in the inequality (11) (the first and the second terms) can be treated in the following way (see Bensoussan [1]). For simplicity, we let

$$f_x(t) = f_x(y(t), u(t)), \quad f_{xx}(t) = f_{xx}(y(t), u(t)) \quad \text{for } f = g, \sigma, l, h.$$

Consider a linear stochastic system

$$\begin{aligned}
&dz(t) = (g_x(t)z(t) + \phi(t))\,dt + (\sigma_x(t)z(t) + \psi(t))\,dB(t), \qquad z(0) = 0, \\
&(\phi(\cdot), \psi(\cdot)) \in L^2_{\mathcal{F}}(0, T; R^n) \times (L^2_{\mathcal{F}}(0, T; R^n))^d, \qquad \Psi = (\psi_1, \cdots, \psi_d),
\end{aligned} \tag{12}$$

where $L^2_{\mathcal{F}}(0, T; R^n)$ is the space of all R^n-valued adapted processes such that

$$E\int_0^T |\phi(t)|^2\,dt<\infty.$$

We can construct a linear functional on the Hilbert space $L^2_{\mathcal{F}}(0, T; R^n)\times (L^2_{\mathcal{F}}(0, T; R^n))^d$ as follows:

$$I(\phi(\cdot), \psi(\cdot)) = E\int_0^T l_x(t)z(t)\,dt + E(h_x(T)z(T)),$$

where $(\phi(\cdot), \psi(\cdot))$ and $z(\cdot)$ are related by (12). It is easy to verify that $I(\cdot,\cdot)$ is continuous. Then by the Riesz Representation Theorem, there is a unique

$$(p(\cdot), K(\cdot)) \in L^2_{\mathcal{F}}(0, T; R^n)\times (L^2_{\mathcal{F}}(0, T; R^n))^d,$$
$$K=(K_1, \cdots, K_d),$$

such that

$$E\int_0^T \left[(p(t), \phi(t)) + \sum_{j=1}^d (K_j(t), \psi_j(t))\right] dt = I(\phi(\cdot), \psi(\cdot)) \tag{13}$$
$$\forall (\phi(\cdot), \psi(\cdot)) \in L^2_{\mathcal{F}}(0, T; R^n)\times (L^2_{\mathcal{F}}(0, T; R^n))^d.$$

With (5) and (6), we can apply this result to some of the terms of (11):

$$E\int_0^T [l_x(s)y_1(s)\,ds + E(h_x(y(T))y_1(T))]$$
$$= E\int_0^T (p(s), g(y(s), u^\varepsilon(s)) - g(y(s), u(s)))\,ds$$
$$+ E\int_0^T \operatorname{tr}[K(s)(\sigma(y(s), u^\varepsilon(s)) - \sigma(y(s), u(s)))]\,ds$$

$$E\int_0^T [l_x(s)y_2(s)\,ds + E(h_x(y(T))y_2(T))$$
$$= E\int_0^T \frac{1}{2}\left[\left(p(s)g_{xx}(s) + \sum_{j=1}^d K_j(s)\sigma^j_{xx}(s)\right) y_1(s)y_1(s)\right] ds$$
$$+ E\int_0^T p^*(s)(g_x(y(s), u^\varepsilon(s)) - g_x(y(s), u(s)))y_1(s)\,ds$$
$$+ E\int_0^T \sum_{j=1}^d K^*_j(s)(\sigma^j_x(y(s), u^\varepsilon(s)) - \sigma^j_x(y(s), u(s)))y_1(s)\,ds.$$

Thus we can rewrite (11) as

$$E\int_0^T (H(y(s), u^\varepsilon(s), p(s), K(s)) - H(y(s), u(s), p(s), K(s)))\,ds$$
$$+\frac{1}{2}E\int_0^T y_1^*(s)H_{xx}(y(s), u(s), p(s), K(s))y_1(s)\,ds \tag{14}$$
$$+\frac{1}{2}Ey_1^*(T)h_{xx}(y(T))y_1(T) \geqq o(\varepsilon),$$

where we denote

$$H(x, v, p, K) = l(x, v) + (p, g(x, v)) + \sum_{j=1}^d (K_j, o^j(x, v)).$$

The interesting thing is that the quadratic terms of (14) can still be treated by applying the Riesz Representation Theorem. Indeed, applying Ito's formula to the matrix-valued processes

$$Y(s)=y_1(s)y_1^*(s)=\begin{bmatrix} y_1^1y_1^1 & \cdots & y_1^1y_1^n \\ \vdots & & \vdots \\ y_1^1y_1^n & \cdots & y_1^ny_1^n \end{bmatrix},$$

we have

$$\begin{aligned} dY(t)=&\left[Y(t)g_x^*(t)+g_x(t)Y(t)+\sum_{j=1}^{d}\sigma_x^j(t)Y(t)\sigma_x^{*j}(t)+\Phi^\varepsilon(t)\right]dt \\ &+[Y(t)\sigma_x^*(t)+\sigma_x(t)Y(t)+\Psi^\varepsilon(t)]\,dB(t), \end{aligned} \tag{15}$$

where

$$\begin{aligned} \Phi^\varepsilon(t)=&\,y_1(t)(g(y(t),u^\varepsilon(t))-g(y(t),u(t)))^*+(g(y(t),u^\varepsilon(t))-g(y(t),u(t)))y_1^*(t) \\ &+\sigma_x(t)y_1(s)(\sigma(y(t),u^\varepsilon(t))-\sigma(y(t),u(t)))^* \\ &+(\sigma(y(t),u^\varepsilon(t))-\sigma(y(t),u(t)))y_1^*(t)\sigma_x^*(t) \\ &+(\sigma(y(t),u^\varepsilon(t))-\sigma(y(t),u(t)))(\sigma(y(t),u^\varepsilon(t))-\sigma(y(t),u(t)))^* \end{aligned}$$

$$\Psi^\varepsilon(t)=y_1(t)(\sigma(y(t),u^\varepsilon(t))-\sigma(y(t),u(t)))^*+(\sigma(y(t),u^\varepsilon(t))-\sigma(y(t),u(t)))y_1^*(t).$$

Consider the following symmetric matrix-valued linear stochastic differential equations:

$$\begin{aligned} dZ(t)=&\left[Z(t)g_x^*(t)+g_x(t)Z(t)+\sum_{j=1}^{d}\sigma_x^j(t)Z(t)\sigma_x^{j*}(t)+\Phi(t)\right]dt \\ &+[Z(t)\sigma_x^*(t)+\sigma_x(t)Z(t)+\Psi(t)]\,dB(t), \end{aligned}$$

$$Z(0)=0,$$

$$(\Phi(\cdot),\Psi(\cdot))\in L^2_{\mathscr{F}}(0,T;R^{n,n})\times(L^2_{\mathscr{F}}(0,T;R^{n,n}))^d,\qquad \Psi=(\psi_1,\psi_2,\cdots,\psi_d),$$

where $R^{n,n}$ is the space of all $n\times n$ real symmetric matrices with the following scalar product:

$$(A_1,A_2)_*=\operatorname{tr}(A_1A_2)\quad \forall A_1,A_2\in R^{n,n}.$$

Now, let us construct a linear functional via (15):

$$M(\Phi(\cdot),\Psi(\cdot))=E\int_0^T(Z(t),H_{xx}(t))_*\,dt+E(Z(T),h_{xx}(y(T)))_*. \tag{16}$$

Obviously, $M(\Phi(\cdot),\Psi(\cdot))$ is a linear continuous functional on

$$L^2_{\mathscr{F}}(0,T;R^{n,n})\times(L^2_{\mathscr{F}}(0,T;R^{n,n}))^d,$$

thus there exists a unique $(P(\cdot),Q(\cdot))\in L^2_{\mathscr{F}}(0,T;R^{n,n})\times(L^2_{\mathscr{F}}(0,T;R^{n,n}))^d$, such that

$$M(\Phi(\cdot),\Psi(\cdot))=E\int_0^T\left[(P(t),\Phi(t))_*+\sum_{j=1}^{d}(Q^j(t),\Psi^j(t))_*\right]dt. \tag{17}$$

Since for all $y\in R^n$, $A\in R^{n,n}$

$$(yy^*,A)_*=\operatorname{tr}[(yy^*)A]=y^*Ay,$$

from (15), (16), (17) we can rewrite (14) as

$$E\int_0^T (H(y(s), u^\varepsilon(s), p(s), K(s)) - H(y(s), u(s), p(s), K(s)))\, ds$$

$$+E\int_0^T \left[(P(t), \Phi^\varepsilon(t))_* + \sum_{j=1}^d (Q_j(t), \Psi_j^\varepsilon(t))_*\right] dt \geqq o(\varepsilon).$$

From the definition of $\Phi^\varepsilon, \Psi^\varepsilon$, we obtain

$$E\int_0^T (H(y(t), u^\varepsilon(t), p(t), K(t)) - H(y(t), u(t), p(t), K(t)))\, dt$$

$$+\frac{1}{2}E\int_0^T \operatorname{tr}[(\sigma(y(t), u^\varepsilon(t)) - \sigma(y(t), u(t)))^* P(s)$$

$$\times(\sigma(y(t), u^\varepsilon(t)) - \sigma(y(t), u(t)))]\, dt \geqq o(\varepsilon).$$

Finally, we have

$$H(y(\tau), v, p(\tau), K(\tau)) - H(y(\tau), u(\tau), p(\tau), K(\tau))$$

$$+\tfrac{1}{2}\operatorname{tr}[(\sigma(y(\tau), v) - \sigma(y(\tau), u(\tau)))^* P(\tau)(\sigma(y(\tau), v) - \sigma(y(\tau), u(\tau)))]$$

$$\geqq 0 \quad \forall v \in U, \quad \text{a.e., a.s.}$$

or, equivalently

$$H(y(\tau), v, p(\tau), K(\tau) - P(\tau)\sigma(y(\tau), u(\tau))) + \tfrac{1}{2}\operatorname{tr}(\sigma\sigma^*(y(\tau), v)P(\tau))$$

$$\geqq H(y(\tau), u(\tau), p(\tau), K(\tau) - P(\tau)\sigma(y(\tau), u(\tau))) \tag{18}$$

$$+\tfrac{1}{2}\operatorname{tr}(\sigma\sigma^*(y(\tau), u(\tau))P(\tau))$$

$$\forall v \in U, \quad \text{a.e., a.s.}$$

Remark. Inequality (18) is the so-called variational inequality (V.I.) of our optimal control problem. In general, it cannot be reduced to the classical form of V.I. except in the case where σ does not depend on the control variable v (see the example at the end of the paper).

5. Adjoint equations and the maximum principle. In this section, we discuss the adjoint equations that solve uniquely the first- and second-order adjoint processes. We give our main result, the maximum principle, at the end of this section. The first-order adjoint equation is the classical one. In fact, from [2] and [3], the first-order adjoint process $(p(\cdot), K(\cdot))$ described in a unique way by (12), (13) is the unique solution of

$$-dp(t) = \left[g_x^*(y(t), u(t))p(t)\right.$$

$$\left. + \sum_{j=1}^d \sigma_x^{j*}(y(t), u(t))K_j(t) + l_x(y(t), u(t))\right] dt - K(t)\, dB(t), \tag{19}$$

$$p(T) = h_x(y(T)).$$

We can also use this result to obtain an equation for $(P(\cdot), Q(\cdot))$. In fact, $(P(\cdot), Q(\cdot))$ is uniquely determined by (16), (17). Thus exactly as in [2] and [3], we can obtain

$$
\begin{aligned}
-dP(t) = \Big[& g_x^*(y(t), u(t))P(t) + P(t)g_x(y(t), u(t)) \\
& + \sum_{j=1}^{d} \sigma_x^{j*}(y(t), u(t))P(t)\sigma_x^j(y(t), u(t)) \\
& + \sum_{j=1}^{d} \sigma_x^{j*}(y(t), u(t))Q_j(t) \\
& + \sum_{j=1}^{d} Q_j(t)\sigma_x^j(y(t), u(t)) + H_{xx}(y(t), u(t), p(t), K(t)) \Big]\, dt \\
& - Q(t)\, dB(t), \\
P(T) = h_{xx}(y(T)).
\end{aligned} \tag{20}
$$

Now we are ready to state our main result.

THEOREM 3. *Let* (3) *hold. If* $(y(\cdot), u(\cdot))$ *is a solution of the optimal control problem* (1), (2), *then we have*

$$(p(\cdot), K(\cdot)) \in L^2_{\mathcal{F}}(0, T; R^n) \times (L^2_{\mathcal{F}}(0, T; R^n))^d,$$

$$(P(\cdot), Q(\cdot)) \in L^2_{\mathcal{F}}(0, T; R^{n,n}) \times (L^2_{\mathcal{F}}(0, T; R^{n,n}))^d,$$

which are, respectively, solutions of (19) *and* (20) *such that the variational inequality* (18) *holds.*

Remark. Recently, we obtained the following result. Under a very heavy additional assumption, we can relate $K(\cdot)$, $P(\cdot)$ by

$$K(t) = P(t)\sigma(y(t), u(t)).$$

So, we can rewrite the maximum principle (18) as

$$
\begin{aligned}
& H(y(\tau), v, p(\tau), 0) + \tfrac{1}{2}\operatorname{tr}(\sigma\sigma^*(y(\tau), v)P(\tau)) \\
& \quad \geqq H(y(\tau), u(\tau), p(\tau), 0) + \tfrac{1}{2}\operatorname{tr}(\sigma\sigma^*(y(\tau), u(\tau))P(\tau)) \quad \forall v \in U, \quad \text{a.e., a.s.}
\end{aligned}
$$

But this relation does not hold in the general case, and not even in the linear quadratic case (see [11]).

6. Problem with final state constraint. We discuss briefly the case when there is an endpoint constraint on the state variable

$$EG(x(T)) = 0,$$

where

$$G(x): R^n \to R^m, \qquad m \leqq n.$$

We also assume that G satisfies the same smoothness condition as h, in (3). We will apply the following well-known Ekeland Lemma (see [5], [6]).

LEMMA 4 (Ekeland's variational principle). *Let* $(V, d(\cdot, \cdot))$ *be a complete metric space, and let*

$$f(\cdot): V \to R$$

be lower-semicontinuous, bounded below. If for all $\rho > 0$ *there exists* $u \in V$ *satisfying*

$$f(u) \leqq \inf_{v \in V} f(v) + \rho,$$

then there exists $u_\rho \in V$, *satisfying the following*:

(i) $f(u_\rho) \leqq f(u)$,

(ii) $d(u, u_\rho) \leqq \rho^{1/2}$,

(iii) $f(v) + \rho^{1/2} d(v, u_\rho) \geqq f(u_\rho)$, *for all* $v \in V$.

We consider only the case where $l(x, v) = 0$ to simplify the statement. Let $(y(\cdot), u(\cdot))$ be an optimal solution of the problem (1), (2) with final state constraint (21). For any $v(\cdot) \in U_{\text{ad}}$, consider the following cost function of the free final system (1):

$$J^\rho(v(\cdot)) = [(Eh(x(T)) - Eh(y(T)) + \rho)^2 + |EG(x(T))|^2]^{1/2}.$$

It is easy to check that

$$0 < J^\rho(v(\cdot)) \qquad \forall v(\cdot) \in U_{\text{ad}},$$

$$J^\rho(u(\cdot)) = \rho.$$

Thus, from Ekeland's variational principle, there is a $v_\rho(\cdot) \in U_{\text{ad}}$, such that

(i) $J^\rho(v_\rho(\cdot)) \leqq \rho$,

(ii) $d(v_\rho(\cdot), u(\cdot)) \leqq \rho^{1/2}$,

(iii) $v_\rho(\cdot)$ is an optimal control of the system (1) under the following cost function:

$$J_d^\rho(v(\cdot)) = J^\rho(v(\cdot)) + \rho^{1/2} d(v(\cdot), v_\rho(\cdot)),$$

where we choose the metric in U_{ad} as

$$d(v(\cdot), u(\cdot)) = E \text{ mes } \{0 \leqq t \leqq T;\ v(t) \neq u(t)\}.$$

It is easy to check that $(U_{\text{ad}}, d(\cdot, \cdot))$ is a complete space. Thus we have an optimal control similar to the one previously discussed (with a slightly different cost function). In fact, we can make a "spike variation" for this optimal control $v_\rho(\cdot)$ as in § 3:

$$v_\rho^\varepsilon(s) = \begin{cases} v & \tau \leqq s \leqq \tau + \varepsilon, \\ v_\rho(s) & \text{otherwise.} \end{cases}$$

Let $x_\rho(\cdot)$, $x_\rho^\varepsilon(\cdot)$ be the trajectories corresponding to $v_\rho(\cdot)$, $v_\rho^\varepsilon(\cdot)$, respectively, and we have

$$|Eh(x_\rho^\varepsilon(T)) - Eh(x_\rho(T)) - Eh_x(x_\rho(T))(y_1(T) + y_2(T)) - \tfrac{1}{2} Eh_{xx}(x_\rho(T)) y_1(T) y_1(T)| \leqq C\varepsilon^2$$

with a constant C independent of ρ, where $y_1(\cdot)$, $y_2(\cdot)$ are solutions of (5), (6) with $y(\cdot) = x_\rho(\cdot)$, $u(\cdot) = v_\rho(\cdot)$. We also have

$$|Eh_x(x_\rho(T))(y_1(T) + y_2(T)) + \tfrac{1}{2} Eh_{xx}(x_\rho(T)) y_1(T) y_1(T)| < C\varepsilon.$$

Similarly,

$$|EG(x_\rho^\varepsilon(T)) - EG(x_\rho(T)) - EG_x(x_\rho(T))(y_1(T) + y_2(T)) - \tfrac{1}{2} EG_{xx}(x_\rho(T)) y_1(T) y_1(T)| \leqq C\varepsilon^2,$$

$$|EG_x(x_\rho(T))(y_1(T) + y_2(T)) + \tfrac{1}{2} EG_{xx}(x_\rho(T)) y_1(T) y_1(T)| \leqq C\varepsilon.$$

Since $J^\rho(v_\rho(\cdot)) > 0$, we have

$$\begin{aligned} 0 &\leqq J^\rho(v_\rho^\varepsilon(\cdot)) + \sqrt{\rho}\, d(v_\rho(\cdot), v_\rho^\varepsilon(\cdot)) - J^\rho(v_\rho(\cdot)) \\ &= J^\rho(v_\rho^\varepsilon(\cdot)) - J^\rho(v_\rho(\cdot)) + \varepsilon\sqrt{\rho} \\ &\leqq \lambda^\rho (Eh_x(x_\rho(T))(y_1(T) + y_2(T)) + \tfrac{1}{2} Eh_{xx}(x_\rho(T)) y_1(T) y_1(T)) \\ &\quad + \mu^\rho (EG_x(x_\rho(T))(y_1(T) + y_2(T)) + \tfrac{1}{2} EG_{xx}(x_\rho(T)) y_1(T) y_1(T)) + \varepsilon\sqrt{\rho} + o(\varepsilon), \end{aligned}$$

where $o(\varepsilon)$ does not depend on ρ:

$$\lambda^\rho = (J^\rho(v_\rho(\cdot)))^{-1/2} E(h(x_\rho(T)) - h(y(T)) + \rho,$$
$$\mu^\rho = (J^\rho(v_\rho(\cdot)))^{-1/2} E(G(x_\rho(T))).$$

Thus, exactly similar as in § 4, we can derive

$$\int_0^T [H(x_\rho(t), v_\rho^\varepsilon(t), p^\rho(t), K^\rho(t) - P^\rho(t)\sigma(x_\rho(t), v_\rho(t))) + \tfrac{1}{2}\operatorname{tr}(\sigma\sigma^*(x_\rho(t), v_\rho^\varepsilon(t))P^\rho(t))]\, dt$$
$$\geqq \int_0^T [H(x_\rho(t), v_\rho(t), p^\rho(t), K^\rho(t) - P^\rho(t)\sigma(x_\rho(t), v_\rho(t))) + \tfrac{1}{2}\operatorname{tr}(\sigma\sigma^*(x_\rho(t), v_\rho(t))P^\rho(t))]\, dt - \varepsilon\sqrt{\rho} - o(\varepsilon),$$

where $p^\rho(\cdot)$, $K^\rho(\cdot)$, $P^\rho(\cdot)$ are the solutions of (see (19), (20))

$$-dp^\rho(t) = \left[g_x^*(x_\rho(t), v_\rho(t))p^\rho(t) + \sum_{j=1}^d \sigma_x^{j*}(x_\rho(t), v_\rho(t))K_j^\rho(t) \right] dt - K^\rho(t)\, dB(t),$$
$$p^\rho(T) = \lambda^\rho h_x(x_\rho(T)) + \mu^\rho G_x(x_\rho(T)),$$
$$-dP^\rho(t) = \Big[g_x^*(x_\rho(t), v_\rho(t))P^\rho(t) + P^\rho(t)g_x(x_\rho(t), v_\rho(t)) + \sigma_x^*(x_\rho(t), v_\rho(t))P^\rho(t)\sigma_x(x_\rho(t), v_\rho(t)) + \sum_{j=1}^d \sigma_x^{j*}(x_\rho(t), v_\rho(t))Q_j^\rho(t) + \sum_{j=1}^d Q_j^\rho(t)\sigma_x^{j*}(x_\rho(t), v_\rho(t)) + H_{xx}(x_\rho(t), v_\rho(t), p^\rho(t), K^\rho(t)) \Big] dt - Q^\rho(t)\, dB(t),$$
$$P(T) = \lambda^\rho h_{xx}(x_\rho(T)) + \mu^\rho G_{xx}(x_\rho(T)).$$

Thus we have

$$H(x_\rho(\tau), v, p^\rho(\tau), K^\rho(\tau) - P^\rho(\tau)\sigma(x_\rho(\tau), v_\rho(\tau))) + \tfrac{1}{2}\operatorname{tr}(\sigma\sigma^*(x_\rho(\tau), v)P^\rho(\tau))$$
$$\geqq H(x_\rho(\tau), v_\rho(\tau), p^\rho(\tau), K^\rho(\tau) - P^\rho(\tau)\sigma(x_\rho(\tau), v_\rho(\tau))) + \tfrac{1}{2}\operatorname{tr}(\sigma\sigma^*(x_\rho(\tau), v_\rho(\tau))P^\rho(\tau)) - \sqrt{\rho} \quad \forall v \in U, \quad \text{a.e., a.s.}$$

Clearly,

$$(\lambda^\rho)^2 + |\mu^\rho|^2 = 1.$$

Thus there exists a subsequence of (λ^ρ, μ^ρ) that converges to a nonzero (λ, μ). On the other hand, by the construct of $v_\rho(\cdot)$, we know that

$$x_\rho(\cdot) \to y(\cdot) \quad \text{in } L_{\mathscr{F}}^2(0, T; R^n).$$

Consequently,

$$p^\rho(\cdot) \to p(\cdot) \quad \text{in } L_{\mathscr{F}}^2(0, T; R^n),$$
$$K^\rho(\cdot) \to K(\cdot) \quad \text{in } (L_{\mathscr{F}}^2(0, T; R^n))^d,$$
$$P^\rho(\cdot) \to P(\cdot) \quad \text{in } L_{\mathscr{F}}^2(0, T; R^{n,n}),$$
$$Q^\rho(\cdot) \to Q(\cdot) \quad \text{in } (L_{\mathscr{F}}^2(0, T; R^{n,n}))^d.$$

Thus we have

$$
\begin{aligned}
&-dp(t)=\left[g_x^*(y(t),u(t))p(t)+\sum_{j=1}^{d}\sigma_x^{j*}(y(t),u(t))K_j(t)\right]dt-K(t)\,dB(t),\\
&p(T)=\lambda h_x(y(T))+\mu G_x(y(T)),
\end{aligned}
\tag{21}
$$

$$
\begin{aligned}
-dP(t)=&\Big[g_x^*(y(t),u(t))P(t)+P(t)g_x(y(t),u(t))\\
&+\sigma_x^*(y(t),u(t))P(t)\sigma_x(y(t),u(t))+\sum_{j=1}^{d}\sigma_x^{j*}(y(t),u(t))Q_j(t)\\
&+\sum_{j=1}^{d}Q_j(t)\sigma_x^{j*}(y(t),u(t))+H_{xx}(y(t),u(t),p(t),K(t))\Big]dt\\
&-Q(t)\,dB(t),\\
P(T)=&\lambda h_{xx}(y(T))+\mu G_{xx}(y(T)),
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
&H(y(\tau),v,p(\tau),K(\tau)-P(\tau)\sigma(y(\tau),u(\tau)))+\tfrac{1}{2}\operatorname{tr}(\sigma\sigma^*(y(\tau),v)P(\tau))\\
&\quad\geqq H(y(\tau),u(\tau),p(\tau),K(\tau)-P(\tau)\sigma(y(\tau),u(\tau)))\\
&\qquad+\tfrac{1}{2}\operatorname{tr}(\sigma\sigma^*(y(\tau),u(\tau))P(\tau))\quad\forall v\in U,\quad\text{a.e., a.s.}
\end{aligned}
\tag{23}
$$

Thus we have Theorem 5.

THEOREM 5. *Let* (3) *hold. If* $(y(\cdot),u(\cdot))$ *is an optimal solution of the optimal control problem* (1), (2) *with final state constraint* (21) *(and* $l(x,v)=0$*), then there are nonzero* $(\lambda,\mu)\in R\times R$

$$(p(\cdot),K(\cdot))\in L(0,T;R^n)\times(L(0,T;R^n))^d,$$

$$(P(\cdot),Q(\cdot))\in L(0,T;R^{n,n})\times(L(0,T;R^{n,n}))^d,$$

that are, respectively, solutions of (21) *and* (22) *such that the variational inequality* (23) *holds.*

To finish this paper, let us consider two special cases.

Case (i). The diffusion coefficient does not contain the control variable $\sigma=\sigma(x)$. In this case (18) becomes

$$H(y(t),v,p(t),K(t))\geqq H(y(t),u(t),p(t),K(t)).$$

This is a well known result (see [5], [6]).

Case (ii). The control domain is convex, and the data are continuously differentiable with respect to v. In this case from (18) we can deduce

$$(H_v(y(t),u(t),p(t),K(t)),v-u(t))\geqq 0\quad\forall v\in U,\quad\text{a.e., a.s.}$$

This coincides with the result of [1] and [2].

Acknowledgments. The author thanks Professor Li Xunjing and all the members of the Control Theory Seminar of Fudan University for useful discussions related to this work. The author also thanks the referees who offered many useful suggestions that improved the first version of the paper.

REFERENCES

[1] A. BENSOUSSAN, *Lecture on stochastic control*, in Nonlinear Filtering and Stochastic Control, Lecture Notes in Mathematics 972, Proc. Cortona, Springer-Verlag, Berlin, New York, 1981.

[2] A. BENSOUSSAN, *Stochastic maximum principle for distributed parameter system*, J. of the Franklin Inst., 315 (1983), pp. 387-406.
[3] ———, *Perturbation Methods in Optimal Control*, Dunod, Gauthier-Villars, 1988.
[4] J. M. BISMUT, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., 20 (1978), pp. 62-78.
[5] I. EKELAND, *Sur les problems variationels*, Acad. Sci. Paris, 275 (1972), pp. 1057-1059.
[6] ———, *Nonconvex minimization problems*, Bull. Amer. Math. Soc. (NS), 1 (1979), pp. 443-474.
[7] U. G. HAUSSMANN, *General necessary conditions for optimal control of stochastic system*, Math. Programming Stud., 6 (1976), pp. 34-48.
[8] Y. HU, *Maximum principle of optimal control for Markov processes*, in Proc. Symposium on Control Theory and Applications, PRC, September, 1987; Acta Math. Sinica, to appear.
[9] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North Holland, Amsterdam, 1981.
[10] H. J. KUSHNER, *Necessary conditions for continuous parameter stochastic optimization problems*, SIAM J. Control, 10 (1972), pp. 550-565.
[11] S. PENG, *Maximum principle for stochastic optimal control with nonconvex domain*, to appear.
[12] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Control Processes*, John Wiley, New York, 1962.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 4, pp. 980-997, July 1990

012

REGULARITY PROPERTIES OF OPTIMAL CONTROLS*

F. H. CLARKE† AND R. B. VINTER‡

Abstract. A class of optimal control problems is considered, involving integral cost functions and linear dynamics. Broad, verifiable hypotheses (HE) are known under which the problems have solutions. These must be supplemented for standard optimality conditions, in the form of a Maximum Principle, to apply. New regularity properties of optimal controls, when merely hypotheses (HE) are in force, are established. A byproduct is a Maximum Principle governing optimal controls under such circumstances, where the costate function is possibly unbounded. Criteria for minimizing controls to be essentially bounded are also deduced.

Key words. optimal controls, bounded controls, regularity, necessary conditions, maximum principle

AMS(MOS) subject classification. 49B05

1. Introduction.

This paper addresses the question of regularity of optimal controls. Except in the special case of optimal control problems that can be reformulated as problems in the calculus of variations, regularity theory has received scant attention as compared, say, with existence theory or necessary conditions of optimality akin to the Maximum Principle. Available regularity results are typically adjuncts to the Maximum Principle; for certain classes of problems the maximization of the Hamiltonian condition directly yields the information that optimal controls must be, say, continuous, or piecewise constant ([10], [13]).

However, regularity theory, in addition to being of interest in its own right, has an important and underestimated bearing on other aspects of the subject. Consider, for example, necessary conditions. We should like to have necessary conditions at our disposal to identify minimizers predicted by existence theory. Yet the hypotheses of existence theory alone (see, e.g., [10]) do not guarantee validity of the standard form of the Maximum Principle, or guarantee even that it makes sense. But, as we show below, the hypotheses of existence theory, supplemented by a priori regularity properties of optimal controls supplied by regularity theory, suffice to extend the applicability of standard necessary conditions and to generate new ones. Thus regularity theory is a source, as well as a byproduct, of necessary conditions.

The optimal control formulation is very broad, and to establish any kind of regularity properties of optimal controls, we need to focus attention on particular classes of problems. The problems considered in this paper are integral cost problems associated with linear time invariant systems:

$$(P)\begin{cases}\text{Minimize } \displaystyle\int_a^b L(t, x(t), u(t))\, dt \\ \text{subject to} \\ \dot{x}(t) = Ax(t) + Bu(t) + d(t), \quad \text{a.e. } t \in [a, b], \\ x(0) = \xi_0,\ x(1) = \xi_1. \end{cases} \tag{1.1}$$

* Received by the editors June 28, 1989; accepted for publication September 18, 1989.

† Centre de recherches mathématiques, Université de Montréal, C.P. 6128-A, Montréal, Québec, H3C 3J7, Canada. This research was supported by the Natural Sciences and Engineering Research Council of Canada. On remercie les Fonds FCAR (Quebec) de leur appui.

‡ Department of Electrical Engineering, Imperial College, Exhibition Road, London SW7 2BT, United Kingdom.

Here a, b are real numbers ($b>a$), $n(>0)$, $m(>0)$ are integers, ξ_0, ξ_1 are n-vectors, and $L:\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$, $d:\mathbb{R}\to\mathbb{R}^n$ given functions.

We term *control* an integrable function $u:[a, b]\to\mathbb{R}$. A solution to the differential equation is called a *state trajectory* (corresponding to u). A state trajectory together with the control to which it corresponds is called a *process.* Precisely stated then, the control problem (P) is to minimize the value of the integral functional over processes, the state trajectories of which take the specified values at times a and b.

Notice that, since we are interested in regularity of optimal controls, we lose nothing by imposing fixed endpoint constraints on the state trajectories. Indeed suppose u^* is an optimal control (with x^* a corresponding state trajectory) for some problem with endpoint constraints $(x(a), x(b))\in C$. Then u^* is certainly still minimizing for the related fixed endpoint problem where we require $x(a)=x^*(a)$, $x(b)=x^*(b)$. Thus regularity results for fixed endpoint problems immediately carry over to problems with any other kind of endpoint constraints.

The hypotheses we impose on the integrand L in the cost function are more or less the weakest which are known to guarantee existence of optimal controls ([10]).

(H1): $L(t, x, u)$ is locally bounded, measurable in t and convex in u. d is an integrable function.

(H2): $L(t, x, u)$ is locally Lipschitz continuous in (x, u) uniformly in t. This means that for each bounded set $S\subset\mathbb{R}^n\times\mathbb{R}^m$, there exists a constant K such that for all $t\in[a, b]$ and (x_1, u_1), $(x_2, u_2)\in S$ the following inequality holds:

$$|L(t, x_1, u_1)-L(t, x_2, u_2)|\leqq K|(x_1-x_2, u_1-u_2)|.$$

(H3): $L(t, x, u)$ is coercive in the following sense. There exists a number $c\geqq 0$ and a convex function $\theta:[0, \infty)\to\mathbb{R}$ such that

$$L(t, x, u)\geqq -c|x|+\theta(|u|)$$

for all $(t, x, u)\in[a, b]\times\mathbb{R}^n\times\mathbb{R}^m$, where $\theta(r)/r\to\infty$ as $r\to\infty$.

These hypotheses will be in force throughout the paper.

Notice that we have not included constraints on the control values in problem (P) and it is possible that the values of optimal controls are unbounded. The Ball-Mizel problem, which may be expressed as an optimal control problem, is one instance where hypotheses (H1)-(H3) are satisfied and such behaviour is observed:

$$P(k, \varepsilon)\begin{cases}\text{Minimize } \displaystyle\int_0^1 [|x^3(t)-t^2|^2|u(t)|^{14}+\varepsilon|u(t)|^2]\,dt\\ \text{subject to}\\ \dot{x}(t)=u,\\ x(0)=0, \qquad x(1)=k.\end{cases}$$

It is known ([1], [5]), that for certain choices of constants k and ε, $P(k, \varepsilon)$ has a unique optimal control $u(t)=kt^{-1/3}$, a function with unbounded values.

Under the circumstances, we should aim to describe the manner in which the optimal controls are unbounded as best we can. Here the notion of "regular point" is helpful.

DEFINITION 1.1. Given a measurable function $g:[0, 1]\to\mathbb{R}^n$ and a point $t\in[a, b]$, t is said to be a regular point of g if there is some neighbourhood of t in $[0, 1]$ on which g is essentially bounded.

The set of regular points is open in $[0, 1]$. It may be empty; g given by

$$g(s)=\sum_i 2^{-i}|s-t_i|^{-1/2},$$

in which $\{t_i\}$ is an ordering of the rationals in $[0, 1]$, is an example of an integrable function with no regular points. In a sense, the larger the set of regular points, the better behaved is the function g. In the extreme case when all points in $[0, 1]$ are regular, the function g is actually essentially bounded.

Our main result is that, for any optimal control, the set of regular points Ω has full measure; we are not able to exclude bad behaviour, of course, but we can at least confine it to a closed set of zero measure (the complement of Ω). This information about regular points leads to new optimality conditions to replace the Maximum Principle when it fails to apply, and to new criteria for optimal controls to be bounded, continuous, etc.

It is true to say that any reasonable design procedure in control engineering, including those based on optimization, must yield controllers with bounded values, because there will always be physical bounds on the outputs of control devices. The problems we study here, while admitting optimal controls with unbounded values, are nonetheless relevant to control engineering. Indeed a common procedure in optimization-based control system design is to replace control constraints, which complicate the analysis, by a penalty term $Q(u)$, for example,

$$Q(u) = \varepsilon \int_a^b |u(t)|^2 \, dt,$$

in the cost. Our theory, among other things, gives criteria for optimal controllers to have bounded values, and therefore has something to say about whether this penalization technique has the desired effect.

Regularity properties of minimizing arcs have long been a field of enquiry in the calculus of variations ([4]-[9], [14]). The present paper builds on this literature, and above all on [9], which treats problems in the calculus of variations with higher order derivatives.

We conclude with some notational points. For integers $\alpha \geqq 2$, $W^{\alpha,1}(I; \mathbb{R}^n)$ is taken to be the space of $(\alpha - 2)$ continuously differentiable n vector-valued functions on the interior of the closed interval I, whose first $(\alpha - 2)$ derivatives have finite limits at the endpoints of I, and whose $(\alpha - 1)$ derivative is absolutely continuous. $W^{1,1}(I; \mathbb{R}^n)$ is the space of n vector valued absolutely continuous functions on I. Given a function y, $D^0 y$, $D^1 y$, $\cdots$ denote the derivatives of y of order $0, 1, \cdots$. For nonnegative integers j and k, with $j \leqq k$, we define $D_j^k y(t) = \operatorname{col}(D^j y(t), D^{j+1} y(t), \cdots, D^k y(t))$. We also employ the notation $\| \ \|_{\infty,s,t}$ for the $L^\infty([s, t])$ norm.

2. The regularity results and a new maximum principle.

THEOREM 2.1. *Suppose there is a process that satisfies the endpoint constraints. Then a minimizing process (x^*, u^*) exists and the set Ω, comprising the regular points of u^*, is open in $[a, b]$ and of full measure. Furthermore, Ω contains all closed subintervals of $[a, b]$, of positive length, on which u^* is (globally) essentially bounded.*

The principal content of the theorem is the assertion that the relatively open set Ω has full measure in $[a, b]$. Included also is a criterion for a point to lie in Ω which may alternatively be expressed: for u^* to be essentially bounded on a relative neighbourhood of a point $t \in [a, b]$, it suffices that u^* be essentially bounded on an interval of positive length immediately to the left, or right, of the point t in question. This criterion will be useful when we seek to show that, in special situations, $\Omega = [a, b]$.

Under additional continuity and strict convexity hypotheses on the data, we shall establish continuity properties of optimal controls and their derivatives on Ω.

THEOREM 2.2. *Let (x^*, u^*) be a minimizing process and let Ω be the set of regular points of u^*. We have*

(i) *if (in addition to* (H1)-(H3)) *for each* t *in* $[a, b]$ *and* $v \in \mathbb{R}^n$ *the function* $w \to L(t, x^*(t), w)$ *is strictly convex and the function* $s \to L(s, x^*(s), v)$ *is continuous at* t, *then* u^* *is continuous on* Ω, *and*

(ii) *if (in addition to* (H1)-(H3) *and the extra conditions in* (i)) *for each* $t \in [a, b]$ *the function* L *is* C^r *in its arguments near* $(t, x^*(t), u^*(t))$ *and* d *is* C^{r-1}, *for some integer* $r \geqq n+1$, *then* u^* *is* C^{r-n} *on* Ω. (*If* Ω *contains an endpoint, a say, then "*u^* *is* C^{r-n} *(at* a*)" is taken to mean that* u^* *is continuous at* a *and, for* $s = 1, \cdots, r-n$, *the derivatives* $D^s u^*$ *exist and are continuous on* $(a, a+\varepsilon)$ *for some* $\varepsilon > 0$, *and* $D^s u^*(t)$ *has a finite limit as* $t \downarrow a$.)

Finally, we give a new version of the Maximum Principle which is valid merely under the hypotheses (H1)-(H3). This involves the (pseudo) Hamiltonian function $H(\cdot, \cdot, \cdot, \cdot)$,

$$H(t, x, u, p) := p \cdot [Ax + Bu] - L(t, x, u).$$

THEOREM 2.3. *Let* (x^*, u^*) *be a minimizing process and let* Ω *be the set of regular points of* u^*. *Then there exists a measurable, row vector valued function* $p : [0, 1] \to \mathbb{R}^n$, *which is locally Lipschitz continuous on* Ω, *and for which*

$$-\dot{p}(t) \in p(t) \cdot A - \partial_x L(t, x^*(t), u^*(t)), \quad a.e., \tag{2.1}$$

and

$$H(t, x^*(t), u^*(t), p(t)) = \max_{u \in \mathbb{R}^m} \{H(t, x^*(t), u, p(t))\}, \quad a.e. \tag{2.2}$$

The novelty of this Maximum Principle is that, in contrast with the standard versions, the costate function p is not required to be absolutely continuous; it may even be unbounded. The costate differential inclusion (2.1) is satisfied in an "almost everywhere" sense and, since there is no guarantee under hypotheses (H1)-(H3) that $t \to \partial_x L(t, x^*(t), u^*(t))$ will have integrable selectors, we cannot necessarily express it as an integral inclusion

$$p(t) \in p(a) + \int_a^t [p(s) \cdot A - \partial_x L(s, x^*(s), u^*(s))]\, ds, \quad \text{for all } t.$$

(Problem $P(k, \varepsilon)$ is one instance where a Maximum Principle of this type is needed; here the costate function corresponding to the optimal control $u(t) = kt^{-1/3}$ is unbounded and fails to satisfy the integral inclusion.) In § 5, we shall apply these results to identify conditions implying that $\Omega = [a, b]$, i.e., that u^* is globally essentially bounded.

3. Proof of the regularity theorems.

3.1. Reduction to a special case. It is convenient at the outset to reduce the optimization problem considered to one where extra hypotheses are assumed to be in force.

(H4): The exogenous term $d(t)$ in the dynamic equations (1.1) is zero.

(H5): The constant c and convex function θ of hypothesis (H3) satisfy $c = 0$ and θ is an increasing function.

(H6): (A, B) is a controllable pair and B has full column rank.

LEMMA. *It suffices to prove Theorems* 2.1 *and* 2.2 *in the special case where, in addition to hypotheses* (H1)-(H3), *we impose* (H4)-(H6).

Proof. We suppose the assertions of Theorems 2.1 and 2.2 are true under hypotheses (H1)-(H6).

Step 1 (Disposal of (H6)). Suppose the data satisfies (H1)-(H5), but possibly not (H6). We treat only the case where (A, B) is not controllable, B does not have full rank and $B \neq 0$ (the other cases one needs to consider "(A, B) is not controllable and B has full rank," etc., are treated by obvious simplifications of the arguments to follow). Changing the basis on the state space (to exhibit the controllable part of the control system and its complement) and the basis on the control space (to exhibit the kernel of B and its complement), yields a nonsingular $n \times n$ matrix P, a nonsingular $m \times m$ matrix Q, and integers $\tilde{n}$, $\tilde{m}$ $(0 < \tilde{n} < n,\ 0 < \tilde{m} < m)$ with the following properties:

$$P^{-1}AP = \tilde{A}, \qquad P^{-1}BQ = \tilde{B}.$$

Here $\tilde{A}$ and $\tilde{B}$ are matrices which may be partitioned thus:

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix}, \qquad \tilde{B} = \begin{bmatrix} 0 & \tilde{B}_{12} \\ 0 & 0 \end{bmatrix}$$

($\tilde{A}_{11}$ is $\tilde{n} \times \tilde{n}$, $\tilde{A}_{12}$ is $\tilde{n} \times (n - \tilde{n})$, $\tilde{A}_{22}$ is $(n - \tilde{n}) \times (n - \tilde{n})$, and $\tilde{B}_{12}$ is $\tilde{n} \times \tilde{m}$). We also have that $(\tilde{A}_{11}, \tilde{B}_{12})$ is a controllable pair and $\tilde{B}_{12}$ has full column rank. These are simple refinements of results proved in [11].

Now define functions $\tilde{x}_2(t)$ and $y(t)$ to satisfy

$$\dot{\tilde{x}}_2(t) = \tilde{A}_{22}\tilde{x}_2(t), \qquad \tilde{x}_2(a) = [P^{-1}\xi_0]_2,$$

and

$$\dot{y}(t) = \tilde{A}_{11}y(t) + \tilde{A}_{12}\tilde{x}_2(t), \qquad y(a) = [P^{-1}\xi_0]_1.$$

Here the notation $[\ ]_1$, $[\ ]_2$ indicates the first and second block components of a partitioned vector, thus

$$P^{-1}\xi_0 = \operatorname{col}\{[P^{-1}\xi_0]_1, [P^{-1}\xi_0]_2\}.$$

Consider now the following optimal control problem (P_1), in which the state vector $\tilde{x}$ is partitioned $\tilde{x} = \operatorname{col}\{\tilde{x}_0, \tilde{x}_1\}$ and the control vector $\tilde{u}$ is partitioned $\tilde{u} = \operatorname{col}\{v, w\}$. Here $\tilde{x}_0$, $\tilde{x}_1$, v, and w are vectors of dimension $m - \tilde{m}$, $\tilde{n}$, $m - \tilde{m}$, and $\tilde{m}$, respectively.

$$\text{Minimize} \int_0^1 \tilde{L}(t, (\tilde{x}_0(t), \tilde{x}_1(t)), (v(t), w(t)))\, dt$$

subject to

$$\frac{d}{dt}\begin{bmatrix} \tilde{x}_0(t) \\ \tilde{x}_1(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{A}_{11} \end{bmatrix}\begin{bmatrix} \tilde{x}_0(t) \\ \tilde{x}_1(t) \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & \tilde{B}_{12} \end{bmatrix}\begin{bmatrix} v(t) \\ w(t) \end{bmatrix},$$

$$\begin{bmatrix} \tilde{x}_0(a) \\ \tilde{x}_1(a) \end{bmatrix} = 0 \quad \text{and} \quad \begin{bmatrix} \tilde{x}_0(b) \\ \tilde{x}_1(b) \end{bmatrix} = \begin{bmatrix} \displaystyle\int_a^b [Q^{-1}u^*(t)]_1\, dt \\ [P^{-1}\xi_1]_1 - y(b) \end{bmatrix}.$$

In this problem the cost integrand $\tilde{L}$ is defined as follows:

$$\tilde{L}(t, (\tilde{x}_0, \tilde{x}_1), (v, w)) := L(t, P \operatorname{col}\{\tilde{x}_1 + y(t), \tilde{x}_2(t)\}, Q \operatorname{col}\{v, w\}).$$

It is a straightforward task to show that

$$\text{(3.1)} \qquad \begin{aligned} &\operatorname{col}\left\{\tilde{x}_0(t) \equiv \int_a^t [Q^{-1}u^*(s)]_1\, ds,\ \tilde{x}_1(t) \equiv [P^{-1}x^*(t)]_1 - y(t)\right\}, \\ &\operatorname{col}\{v(t), w(t)\} \equiv Q^{-1}u^*(t) \end{aligned}$$

solves this problem.

Now we observe that the data for problem (P_1) satisfies hypotheses (H1)-(H5) for appropriately chosen θ and c. However, (H6) is also satisfied. We know then that the process for (P_1) given by (3.1) has the properties listed in Theorems 2.1 and 2.2. But the same is then true of the solution (x^*, u^*) to (P) since it is expressible in terms of the solution (3.1) to (P_1) according to

$$x^*(t) \equiv P \operatorname{col}\{\tilde{x}_1(t) + y(t), \tilde{x}_2(t)\}, \qquad u^*(t) = Q \operatorname{col}\{v(t), w(t)\}.$$

We conclude that the assertions of Theorems 2.1 and 2.2 are true merely under hypotheses (H1)-(H5).

Step 2 (Disposal of (H5)). Next assume (H1)-(H4) are satisfied, but possibly not (H5). Consider $\tilde{\theta}: [0, \infty) \to \mathbb{R}$ defined by

$$\tilde{\theta}(r) := \inf\{\theta(r'): r' > r\}.$$

Since $\tilde{\theta}$ is superlinear and convex, and is majorized by θ, (H3) remains in force when $\tilde{\theta}$ replaces θ. Note that $\tilde{\theta}$ is an increasing function.

Take k to be a constant such that $\|x^*\|_{\infty,a,b} < k$. Consider a new problem (P_2) in which the Lagrangian L of problem (P) is replaced by $L_2(t, y, w) := \max[L(t, y, w), -ck + \theta(|w|)] + ck$. We easily check that (H1)-(H4), and also (H5), are satisfied by (P_2) (with $\theta = \tilde{\theta}$ and $c = 0$). We have that $L_2 \geqq L + ck$ everywhere and $L_2(t, y, w) = L(t, y, w) + ck$ for all points w and all points (t, y) in some tube about x^*. Consequently, (x^*, u^*) is a minimizing process for (P_2) also. Applying the special cases of Theorems 2.1 and 2.2 to (P_2) where (H1)-(H5) are in force, we deduce the assertions of these theorems for (P), merely under (H1)-(H4).

Step 3 (Disposal of (H4)). Finally assume the data of (P) satisfy (H1)-(H3). Consider the problem (P_3)

$$\begin{cases} \text{Minimize } \displaystyle\int_a^b L_3 \, dt \\ \text{subject to } \dot{x} = Ax + Bu \quad \text{and} \\ x(a) = \xi_0, \qquad x(b) = \xi_1'. \end{cases}$$

Here

$$L_3(t, y, w) := L(t, g(t) + x(t), w)$$

and $\xi_1' := \xi_1 - g(b)$, where the function g is the solution to the differential equation

$$\dot{g} = Ag + d, \qquad g(a) = 0.$$

Evidently the process $(x^*(t) - g(t), u^*(t))$ is minimizing for (P_3), and the data for (P_3) satisfy hypotheses (H1)-(H4). Applying Theorems 2.1 and 2.2 to (P_3), as is permissible since (H1)-(H4) are in force, we verify the assertions of Theorems 2.1 and 2.2, in relation to problem (P), when merely (H1)-(H3) are in force.

We are justified then in adding (H4)-(H6) to the list of hypotheses.

3.2. Canonical representation of the dynamic equations. Our purpose here is to show that an arbitrary process for problem (P) can be expressed in terms of higher order derivatives of a single vector valued function. This will pave the way to reducing problem (P) to a problem in the calculus of variations with higher order derivatives, the minimizers for which have known regularity properties.

The required relationships are summarized in the following proposition (whose proof is to be found in the Appendix). Recall that the supplementary hypotheses

(H4)-(H6) are in force and so, in particular, the proposition deals with the situation in which the dynamic equations take the form

$$\dot{x} = Ax + Bu,$$

where (A, B) is a controllable pair and B has full column rank.

PROPOSITION 3.1. *There exists a positive integer α, an αm column vector q, and matrices F (det $F \neq 0$), G and H of dimension $m \times m$, $m \times (\alpha m)$, and $n \times (\alpha m)$, respectively, with the following properties:*

Define the mapping

$$Z : \mathcal{S} \to \tilde{\mathcal{S}},$$

$$\mathcal{S} := \{\text{processes } (x, u) \text{ for (1.1): } x(a) = \xi_0\},$$

$$\tilde{\mathcal{S}} := \{z \in W^{\alpha,1}([a, b]; \mathbb{R}^m) : D_0^{\alpha-1} z(a) = q\},$$

to take the value z at $(x, u) \in \mathcal{S}$, where z is the unique solution to the higher order differential equation

$$\begin{cases} FD^{\alpha} z(t) + GD_0^{\alpha-1} z(t) = u(t), & \text{a.e. } t \in [a, b], \\ D_0^{\alpha-1} z(a) = q. \end{cases}$$

Then Z is a bijection and if $z = Z(x, u)$, we have

$$x(t) = HD_0^{\alpha-1} z(t), \quad \text{for all } t \in [a, b].$$

Furthermore, for any open subinterval $I \subset [a, b]$, we have

(a) *$u \in L^{\infty}(I)$ if and only if $D^{\alpha} z \in L^{\infty}(I)$ and*

(b) *for any integer $\beta \geqq 0$, $u \in C^{\beta}(I)$ if and only if $D^{\alpha} z \in C^{\beta}(I)$.*

3.3. Variational problems with higher order derivatives. This subsection provides a summary for future reference of known regularity properties of minimizers for problems in the calculus of variations involving higher order derivatives. Consider

$$(Q) \begin{cases} \text{Minimize } \displaystyle\int_a^b M(t, D_0^{\alpha-1} z(t), D^{\alpha} z(t))\, dt \\ \text{over arcs } z \in W^{\alpha,1}([a, b]; \mathbb{R}^{\beta}) \text{ such that} \\ D_0^{\alpha-1} z(a) = \xi_0 \quad \text{and} \quad D_0^{\alpha-1} z(b) = \xi_1. \end{cases}$$

Here a, b are given real numbers $(a > b)$. α and β are positive integers. ξ_0 and ξ_1 are given $\alpha\beta$ vectors, and $M : \mathbb{R} \times \mathbb{R}^{\alpha\beta} \times \mathbb{R}^{\beta} \to \mathbb{R}$ is a given function. We shall invoke the following hypotheses:

(I1): $M(t, z, w)$ is locally bounded, measurable in t and convex in w.

(I2): $M(t, z, w)$ is locally Lipschitz continuous in (z, w), uniformly in t.

(I3): There exists $\delta \geqq 0$ and a convex function $\gamma : [0, \infty) \to \mathbb{R}$ such that

$$M(t, z, w) \geqq -\delta|z| + \gamma(|w|),$$

for all $(t, z, w) \in [a, b] \times \mathbb{R}^{\alpha\beta} \times \mathbb{R}^{\beta}$, where $\gamma(r)/r \to \infty$ as $r \to \infty$.

THEOREM 3.2. *Assume* (I1)-(I3). *A minimizing arc z for problem (Q) exists. Let J be the regular points of $D^{\alpha} z$. Then J is an open set of full measure and J contains all closed subintervals of $[a, b]$ on which $D^{\alpha} z$ is essentially bounded. Furthermore*

(i) *If in addition to* (I1)-(I3) *we assume that, for each* t *in* $[a, b]$ *and* v *in* $\mathbb{R}^\beta$, *the function* $w \to M(t, D_0^{\alpha-1}z(t), w)$ *is strictly convex and the function* $s \to M(s, D_0^{\alpha-1}z(s), v)$ *is continuous at* t, *then* $D^\alpha z$ *is continuous on* J, *and*

(ii) *If in addition to the hypotheses of* (i) *we assume that, for each* $t \in [a, b]$, *the function* M *is* C^r *near* $(t, D_0^{\alpha-1}z(t), D^\alpha z(t))$ *for some* $r \geqq \alpha + 1$ *and* $M_{ww}(t, D_0^{\alpha-1}z(t), D^\alpha z(t)) > 0$, *then* $D^\alpha z$ *is* $C^{r-\alpha}$ *on* J.

This theorem is a recasting of [9, Thm. 2.1], suitable for present applications. As an aid to relating the two, we remark that the set of points which are regular points of z in the sense of [9, Thm. 2.1] coincides with the regular points of $D^\alpha z$ in the sense of this paper, and contains all closed subintervals of $[a, b]$, of positive length, on which $D^\alpha z$ is essentially bounded.

3.4. Construction of the auxiliary Lagrangian. Our task here is to associate with the minimizing process (x^*, u^*) a minimizer for a problem of the kind treated in § 3.3. Reference will be made to the mapping Z of Proposition 3.1.

Define

$$z^* \in W^{\alpha,1}([a, b]; \mathbb{R}^m) \quad \text{according to}$$

$$z^* := Z(x^*, u^*)$$

and set

$$K := \|D_0^{\alpha-1}z^*\|_{\infty,a,b}.$$

Now define the function $\varphi : \mathbb{R}^m \to \mathbb{R}$ to be

$$\varphi(w) := \inf\{L(t, Hy, Fw + Gy) : t \in [a, b], |y| \leqq K + 1\}.$$

Define also the real value functions $\tilde{\tilde{L}}$ and $\tilde{L}$ on $\mathbb{R} \times \mathbb{R}^{\alpha m} \times \mathbb{R}^m$ to be

$$\tilde{\tilde{L}}(t, y, w) := L(t, Hy, Fw + Gy)$$

and

$$\tilde{L}(t, y, w) := \max\{\tilde{\tilde{L}}(t, y, w), \operatorname{co} \varphi(w)\}.$$

Here F, G, and H are the matrices of Proposition 3.1. co φ denotes the function whose epigraph is the convex closure of epi φ.

We list important properties of the integrand $\tilde{L}$ just constructed.

PROPOSITION 3.2.

(a) $\tilde{L}$ *satisfies the hypotheses* (I1)-(I3) *of* § 3.2.

(b) $\tilde{L}(t, y, w) \geqq L(t, Hy, Fw + Gy)$ *for all* $t \in [a, b]$, $y \in \mathbb{R}^{\alpha m}$ *and* $w \in \mathbb{R}^m$.

(c) *For all* $t \in [a, b]$, $y \in \mathbb{R}^{\alpha m}$ *such that* $|y| \leqq K + 1$ *and* $w \in \mathbb{R}^m$ *we have*

$$\tilde{L}(t, y, w) = L(t, Hy, Fw + Gy).$$

Proof. By (H5), $\theta(\cdot)$ is a nondecreasing function. It follows that φ is bounded below by $\theta(0)$. Consequently we know that co φ is a finite valued convex function on $\mathbb{R}^m$; it is therefore locally Lipschitz continuous by the properties of convex functions. In particular, φ is locally bounded.

$\tilde{L}(t, y, w)$ is locally bounded since it is the pointwise maximum of two locally bounded functions. For fixed y, w, $\tilde{L}(\cdot, y, w)$ is the pointwise maximum of the measurable function $L(\cdot, Hy, Fw + Gy)$ and the constant function with value co φ, and is therefore measurable. For fixed t, y the function $\tilde{L}(t, y, \cdot)$ is convex, since it is expressible as the pointwise supremum of convex functions. Hypothesis (I1) is verified.

That $\tilde{L}(t, \cdot, \cdot)$ is locally Lipschitz continuous, uniformly in t, follows from the fact that it is the pointwise supremum of two functions $\tilde{\tilde{L}}$ and φ having this property. This is hypothesis (I2).

Define the numbers $c_1(>0)$ and $c_2(\geqq 0)$ to be $c_1 := \|F^{-1}\|$ and $c_2 := \|F^{-1}G\|(K+1)$. ($\|\ \|$ denotes the operator norm.) Since θ is nondecreasing and convex (hypothesis (H5)), we have $\operatorname{co} \varphi(w) \geqq \theta'(|w|)$, for all $w \in \mathbb{R}^n$, where θ' is the convex, finite valued function

$$\theta'(r) := \theta(\max[0, c_1 r - c_2]), \quad \text{for } r \geqq 0.$$

But θ is superlinear, so

$$\lim_{r\to\infty} \frac{\theta'(r)}{r} = \lim_{r\to\infty} \frac{\theta(c_1 r - c_2)}{(c_1 r - c_2)} \times \frac{c_1 r - c_2}{r} = +\infty,$$

i.e., θ' is also superlinear. We have

$$\tilde{L}(t, y, w) \geqq \operatorname{co} \varphi(w) \geqq \theta'(|w|).$$

Hypothesis (I3) is thereby verified.

Properties (b) and (c) follow readily from the definitions of $\tilde{L}$, $\tilde{\tilde{L}}$, and φ.

3.5. Conclusion of proof. Once again take $z^* = Z(x^*, u^*)$ where (x^*, u^*) is the minimizing process for (P) of interest. z^* satisfies the boundary condition $D_0^{\alpha-1} z^*(a) = q$ on elements in $\tilde{\mathcal{F}}$. Define $r := D_0^{\alpha-1} z^*(b)$. In what follows, $\tilde{L}$ is the Lagrangian defined in § 3.4.

LEMMA 3.4. *The element z^* is a minimizer of the functional $\tilde{J}$,*

$$\tilde{J}(z) := \int_a^b \tilde{L}(t, D_0^{\alpha-1} z(t), D^\alpha z(t))\, dt$$

over arcs $z \in W^{\alpha,1}([a, b]; \mathbb{R}^m)$ which satisfy

$$D_0^{\alpha-1} z(a) = q \quad \text{and} \quad D_0^{\alpha-1} z(b) = r.$$

Proof. Take an arbitrary element in $W^{\alpha,1}$ such that

$$D_0^{\alpha-1} z(a) = D_0^{\alpha-1} z^*(a) \quad \text{and} \quad D_0^{\alpha-1} z(b) = D_0^{\alpha-1} z^*(b).$$

Let $(x, u) = Z^{-1}(z)$. By Proposition 3.1,

$$x(a) = HD_0^{\alpha-1} z(a) = HD_0^{\alpha-1} z^*(a) = x^*(a) = \xi_0.$$

Likewise $x(b) = \xi_1$. Thus the process (x, u) satisfies the endpoint constraints on problem (P).

We have

$$\begin{aligned} \tilde{J}(z) &= \int_a^b \tilde{L}(t, D_0^{\alpha-1} z, D^\alpha z)\, dt \\ &\geqq \int_a^b L(t, HD_0^{\alpha-1} z, FD^\alpha z + GD_0^{\alpha-1} z)\, dt = \int_a^b L(t, x, u)\, dt, \end{aligned}$$

by Proposition 3.1 and Proposition 3.2(b),

$$\geqq \int_a^b L(t, x^*, u^*)\, dt,$$

since (x^*, u^*) is minimizing for (P),

$$= \int_a^b L(t, HD_0^{\alpha-1}z^*, FD^\alpha z^* + GD_0^{\alpha-1}z^*)\, dt$$

$$= \int_a^b \tilde{L}(t, D_0^{\alpha-1}z^*, D^\alpha z^*)\, dt = \tilde{J}(z^*),$$

in view of the fact that $\|D_0^{\alpha-1}z^*\| \leqq K$ and by Proposition 3.1 and Proposition 3.2(c). This establishes that z^* is a minimizer.

The requisite hypotheses are satisfied for application of Theorem 3.2 to the problem of minimizing $\tilde{J}$, with reference to the minimizer z^*. We conclude existence of an open set $\Omega \subset [a, b]$ of full measure, on which $D^\alpha z^*$ is locally essentially bounded. But then u^* is also locally essentially bounded on Ω by Proposition 3.1. Theorem 2.1 is proved.

We turn now to Theorem 2.2. Suppose that, in addition to (H1)-(H3), $(t, v) \to L(t, x^*(t), v)$ is continuous in t and strictly convex in v. Define $\xi(t, w) := \tilde{L}(t, D_0^{\alpha-1}z^*(t), w)$. Since $\|D_0^{\alpha-1}z^*\|_{\infty,a,b} < K$, we have from Proposition 3.2 and Proposition 3.1

$$\xi(t, w) = L(t, x^*(t), Fw + GD_0^{\alpha-1}z^*(t)).$$

In view of the continuity of $t \to L(t, x^*(t), v)$ for each v, the continuity of $t \to D_0^{\alpha-1}z^*(t)$ and the fact that $L(t, y, w)$ is locally Lipschitz continuous in y, w, uniformly in t, the function $\xi(t, w)$ is continuous in t for each w. It is strictly convex in w for each t since $L(t, x^*(t), \cdot)$ is strictly convex and F is nonsingular. Theorem 3.2 now tells us that $D^\alpha z^*$ is continuous on Ω. By Proposition 3.1, u^*, too is continuous on Ω.

Finally, we suppose that the hypotheses of Theorem 2.2(ii) are in force. Then for each $t \in [a, b]$, $L(t, y, w)$ is C^r near $(t, x^*(t), u^*(t))$. By Proposition 3.2, there exists $\varepsilon > 0$ such that

$$\tilde{L}(t, y, w) = L(t, Hy, Fw + Gy),$$

for all $t \in [a, b]$, $y \in D_0^{\alpha-1}z^*(t) + \varepsilon B$ and $w \in \mathbb{R}^n$. Another way of expressing this relationship is that

$$\tilde{L}(t, y, w) = L \circ \chi(t, y, u),$$

for all (t, y) lying in a tube about graph $\{D_0^{\alpha-1}z^*\}$ and all $w \in \mathbb{R}^m$. Here χ is the linear map

$$(t, y, w) \xrightarrow{\chi} (t, Hy, Fw + Gy).$$

Now

$$\chi(t, D_0^{\alpha-1}z^*(t), D^\alpha x^*(t)) = (t, x^*(t), u^*(t)),$$

for all $t \in [a, b]$. For each $t \in [a, b]$ then, $\tilde{L}$ is C^r near $(t, D_0^{\alpha-1}z^*(t), D^\alpha z^*(t))$ since it is a composition of maps χ and L which are C^r on neighbourhoods of $(t, D_0^{\alpha-1}z^*(t), D^\alpha z^*(t))$ and $\chi(t, D_0^{\alpha-1}z^*(t), D^\alpha z^*(t))$, respectively. We have already shown that if the data for problem (P) satisfies the extra hypotheses of Theorem 2.2, part (i), at (x^*, u^*), then the data of problem $(\tilde{P})$ satisfies the extra hypotheses of Theorem 3.2 at z^* $(= Z(x^*, u^*))$. We are justified then in applying Theorem 3.2, part (ii), to problem $(\tilde{P})$ at z^*. This tells us that $D^\alpha z^*$ is C^r on Ω. By Proposition 3.1, u^* is also C^r on Ω. This concludes the proof.

4. Proof of Theorem 2.3. Assume that (A, B) is controllable (we shall relinquish this extra hypothesis later). The set Ω is open in $[a, b]$. It is therefore expressible as the countable union of disjoint, relatively open intervals in $[a, b]$,

$$\Omega = \bigcup_{i=1}^{\infty} J_i.$$

Denote by α, β the left and right endpoints of J_1 and take $\tau \in (\alpha, \beta)$. Let $\{\sigma_i\}$ and $\{\tau_i\}$ be monotone sequences in (α, τ) and (τ, β), respectively, such that $\sigma_i \downarrow \alpha$ and $\tau_i \uparrow \beta$ and $i \to \infty$.

For each j, the control process on $[\sigma_j, \tau_j]$, obtained by restricting (x^*, u^*) to $[\sigma_j, \tau_j]$, minimizes $\int L\,dt$ subject to the appropriate fixed endpoint conditions. Also u^* is essentially bounded on $[\sigma_j, \tau_j]$ by Theorem 2.1. This essential boundedness property ensures that the hypotheses are satisfied for application of the Maximum Principle ([3, Thm. 5.2.1]). We deduce existence of Lipschitz continuous functions $q_j : [\sigma_j, \tau_j] \to \mathbb{R}^n$ such that

$$-\dot{q}_j(t) \in q_j(t) \cdot A - \partial_x L(t, x^*(t), u^*(t)), \quad \text{a.e. } [\sigma_j, \tau_j] \tag{4.1}$$

and

$$q_j(t) \cdot Bu^*(t) - L(t, x^*(t), u^*(t)) = \max_w \{q_j(t) \cdot Bw - L(t, x^*(t), w)\}, \quad \text{a.e. } [\sigma_j, \tau_j]. \tag{4.2}$$

(Notice that, since (A, B) is assumed controllable, the Maximum Principle applies in "normal form.") The last equation implies

$$q_j(t) \cdot B \in \partial_u L(t, x^*(t), u^*(t)), \quad \text{a.e. } [\sigma_j, \tau_j]. \tag{4.3}$$

We deduce from (4.1) and (4.3) existence of essentially bounded, measurable functions $\xi_j : [\sigma_j, \tau_j] \to \mathbb{R}^n$ and $\eta_j : [\sigma_j, \tau_j] \to \mathbb{R}^m$ such that, for $j = 1, 2, \cdots$

$$-\dot{q}_j(t) = q_j(t) \cdot A - \xi_j(t), \quad \text{a.e. } [\sigma_j, \tau_j] \tag{4.4}$$

and

$$q_j(t) \cdot B = \eta_j(t), \quad \text{a.e. } [\sigma_j, \tau_j]. \tag{4.5}$$

Furthermore, there exists an increasing sequence of numbers $\{\bar{c}_k\}$ such that, for fixed k,

$$\|\xi_j(\cdot)\|_{\infty, \sigma_k, \tau_k}, \|\eta_j(\cdot)\|_{\infty, \sigma_k, \tau_k} \leqq \bar{c}_k, \tag{4.6}$$

for all $j \geqq k$.

We show that there is a uniform bound on $\|q_j(\tau)\|$, for $j = 1, 2, \cdots$. From (4.4) and the variation of constants formula it follows that

$$q_j(t) = q_j(\tau) \cdot e^{-A(t-\tau)} + s_j(t), \quad \text{a.e. } [\sigma_1, \tau_1], \tag{4.7}$$

for $j = 1, 2, \cdots$, where

$$s_j(t) = \int_\tau^t \xi_j(s)\, e^{-A(t-s)}\, ds.$$

Postmultiplying across equation (4.7) by $BD(t)$, where

$$D(t) := B'\, e^{-A(t-\tau)}\, ds,$$

integrating over $[\tau, \tau_1]$ and noting (4.5), we obtain

$$q_j(\tau) \cdot W(\tau, \tau_1) = \int_\tau^{\tau_1} [\eta_j(t) - s_j(t)B] D(t)\, dt,$$

where $W(\tau, \tau_1)$ is the controllability Grammian of $(-A, B)$, namely,

$$W(\tau, \tau_1) := \int_{\tau}^{\tau_1} e^{-As} BB' e^{-A's}\, ds.$$

Since (A, B) is controllable, $W(\tau, \tau_1)$ is invertible. The uniform boundedness of $\xi_j(\cdot)$, $\eta_j(\cdot)$, and hence of $s_j(\cdot)$, on $[\tau, \tau_1]$ (see (4.6)) ensures existence of a number d such that

$$\|q_j(\tau)\| \leqq d, \quad \text{for } j = 1, 2, \cdots.$$

This establishes the uniform bound on the $\|q_j(\tau)\|$'s.

It follows from this property, (4.4), and (4.6), via application of Gronwall's inequality, that there exist numbers c_k, $k = 1, 2, \cdots$, such that for each k, and $j = k, k+1, \cdots$

$$\|q_j(\cdot)\|_{\infty, \sigma_k, \tau_k} \leqq c_k.$$

We conclude from (4.4) that for each k, the functions $q_j(\cdot), j = k, k+1, \cdots$ are uniformly bounded and equicontinuous on the fixed subinterval $[\sigma_k, \tau_k]$.

Ascoli's theorem now permits us to construct a nested family of subsequences of $\{q_j(\cdot)\}$,

$$\{q_{1j}(\cdot)\}_j \supset \{q_{2j}(\cdot)\}_j \supset \cdots,$$

with the following properties: for fixed k, terms in the sequence $\{q_{kj}\}_{j=1}^{\infty}$ are selected from $\{q_j(\cdot)\}_{j=k}^{\infty}$ and

$$q_{kj}(t) \to p_k(t), \quad \text{uniformly on } [\sigma_k, \tau_k] \tag{4.8}$$

for some continuous function $p_k(\cdot)$.

Take k an arbitrary index value. Then since $\{q_{kj}\}_{j=1}^{\infty}$ is a subsequence of $\{q_j\}_{j=k}^{\infty}$, (4.1) and (4.2) remain valid when q_{kj} replaces q_j and the fixed time interval $[\sigma_k, \tau_k]$ replaces $[\sigma_j, \tau_j]$, for $j = 1, 2, \cdots$. In view of (4.8), standard arguments (see, e.g., [3, p. 201 et seq.]) permit us to pass to the limit in these relationships and obtain the following information about $p_k(\cdot)$: p_k is a Lipschitz continuous function on $[\sigma_k, \tau_k]$,

$$-\dot{p}_k(t) \in p_k(t) \cdot A - \partial_x L(t, x^*(t), u^*(t)), \quad \text{a.e. } [\sigma_k, \tau_k] \tag{4.9}$$

and

$$p_k(t) \cdot Bu^*(t) - L(t, x^*(t), u^*(t)) = \max_w \{p_k(t) \cdot Bw - L(t, x^*(t), w)\}, \tag{4.10}$$

$$\text{a.e. } [\sigma_k, \tau_k].$$

Now the fact that the sequences $\{q_{kj}\}_{j=1}^{\infty}$, $k = 1, 2, \cdots$ are nested ensures that

$$p_j(t) = p_k(t), \quad \text{all } t \in [\sigma_j, \tau_j],$$

for any index values i, j, k such that $i \leqq j \leqq k$.

We may define a function $p : J_1 \to \mathbb{R}^n$ then in the following manner: if $t \in J_1$

$$p(t) = p_j(t),$$

where j is any index value such that $t \in [\sigma_j, \tau_j]$. Since each of the $p_j(\cdot)$'s are Lipschitz continuous, $p(\cdot)$ is locally Lipschitz continuous on J_1.

Let $I_k \subset [\sigma_k, \tau_k]$ be the nullset of points at which either (4.9) or (4.10) is not satisfied. Evidently $p(\cdot)$ satisfies (4.9) and (4.10) for all $t \in J_1 \backslash \cup_i I_k$, a subset of J_1 having full measure.

We now extend $p(\cdot)$ to all of $[0, 1]$. It is constructed on $J_2 \cup J_3 \cup \cdots$ precisely as on J_1. Values are arbitrarily assigned on the nullset $[0, 1] \setminus \cup_i J_i$. Retaining the symbol $p(\cdot)$ for the extension, we see that $p(\cdot)$ is locally Lipschitz continuous on J and satisfies (2.1) and (2.2) almost everywhere. This completes proof of the theorem in the case where (A, B) is controllable.

Suppose now that (A, B) is not controllable. Consider first the case $B = 0$. Then $u^*(\cdot)$ must satisfy

$$L(t, x^*(t), u^*(t)) = \min_u \{L(t, x^*(t), u)\}, \quad \text{a.e.}$$

We readily deduce from this property and hypotheses (H1) and (H3) that $u^*(\cdot)$ is essentially bounded on $[0, 1]$. But then the theorem is just a special case of the standard Maximum Principle [3, Thm. 5.1.2].

If $B \neq 0$, we can find a nonsingular $n \times n$ matrix P and an integer n_1, $0 < n_1 < n$, such that the matrices $\tilde{A}$ and $\tilde{B}$,

$$\tilde{A} := P^{-1}AP \quad \text{and} \quad \tilde{B} = P^{-1}B$$

have structure

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix}, \qquad \tilde{B} = \begin{bmatrix} \tilde{B}_1 \\ 0 \end{bmatrix}.$$

Here $\tilde{A}_{11}$, $\tilde{A}_{12}$, $\tilde{A}_{22}$, and $\tilde{B}_1$ have dimensions $n_1 \times n_1$, $n_1 \times (n - n_1)$, $(n - n_1) \times (n - n_1)$, and $n_1 \times m$, respectively. Furthermore $(\tilde{A}_{11}, \tilde{B})$ is controllable.

By considering a change of coordinates on the state space, $\operatorname{col}\{\tilde{x}_1, \tilde{x}_2\} = P^{-1}x$, we can easily see that the process $(\tilde{x}_1^*, u^*)$ solves the problem,

$$\text{Minimize} \int_0^1 L(t, P_1\tilde{x}_1(t) + P_2\tilde{x}_2^*(t), u(t))\, dt$$

$$\text{over processes} \quad (\tilde{x}_1, u) \quad \text{such that}$$

$$\dot{\tilde{x}}_1(t) = \tilde{A}_{11}\tilde{x}_1(t) + \tilde{B}_1 u(t) + [P^{-1}d(t)]_1,$$

$$\tilde{x}(0) = [P^{-1}\xi_0]_1 \quad \text{and} \quad \tilde{x}(1) = [P^{-1}\xi_1]_1.$$

Here $[P_1, P_2]$ is a partitioning of P into matrices of n_1 and $n - n_1$ columns, respectively, $[\cdots]_1$ ($[\cdots]_2$) denote vectors formed of the first n_1 (last $(n - n_1)$) entries of a given n vector and $(\tilde{x}_1^*(t), \tilde{x}_2^*(t)) = P^{-1}x^*(t)$.

The optimization problem arrived at satisfies the hypotheses of Theorem 2.3 and, in addition, has controllable dynamics. Applying then a special case of Theorem 2.3 that we have proved, we obtain a function $\tilde{p}_1 : [0, 1] \to \mathbb{R}^n$ which is locally Lipschitz continuous on j and satisfies

$$-\dot{\tilde{p}}_1 \in \tilde{p}_1 \cdot \tilde{A}_{11} - \partial_x L(t, P_1\tilde{x}_1^*(t) + P_2\tilde{x}_2^*(t), u^*(t))P_1, \quad \text{a.e.} \tag{4.11}$$

and

$$\begin{aligned} &\tilde{p}_1 \cdot \tilde{B}_1 u^*(t) - L(t, P_1\tilde{x}_1^*(t) + P_2\tilde{x}_2^*(t), u^*(t)) \\ &\qquad = \max_u \{\tilde{p}_1 \cdot \tilde{B}_1 u - L(t, P_1\tilde{x}_1^*(t) + P_2\tilde{x}_2^*(t), u)\}, \quad \text{a.e.} \end{aligned} \tag{4.12}$$

Next, a simpler version of a technique already employed in this proof is used to construct a function $\tilde{p}_2 : [0, 1] \to \mathbb{R}^{n-n_1}$ which is locally Lipschitz continuous on J and satisfies

$$-\dot{\tilde{p}}_2(t) \in \tilde{p}_2(t) \cdot \tilde{A}_{22} + \tilde{p}_1(t) \cdot A_{12} - \partial_x L(t, x^*(t), u^*(t))P_2, \quad \text{a.e.} \tag{4.13}$$

Briefly, we construct the values of $\tilde{p}_2$ successively on $J_1, J_2, \cdots$ (where $J = \cup_j J_i$ is a decomposition of J into relatively open, disjoint intervals). It suffices to say that, in place of an application of the standard Maximum Principle, to assure existence of a Lipschitz continuous solution on a closed subinterval of J to the relevant costate differential equation, we need only appeal to existence theorems for solutions to ordinary differential equations (involving arbitrary selectors of $\partial_x L(t, x^*, u^*)P_2$). Also, instead of using a "controllability" argument to obtain bounds on these solutions at a point in each of the J_j's, we simply set the value to zero there.

Noting $P_1\tilde{x}_1^* + P_2\tilde{x}_2^* = x^*$, we assemble (4.11), (4.12), and (4.13) to obtain

$$-\frac{d}{dt}(\tilde{p}_1, \tilde{p}_2) \in (\tilde{p}_1, \tilde{p}_2)\tilde{A} - \partial_x LP, \quad \text{a.e.} \tag{4.14}$$

and

$$(\tilde{p}_1, \tilde{p}_2)\tilde{B}u^* - L(t, x^*, u^*) = \max_u \{(\tilde{p}_1, \tilde{p}_2)\tilde{B}u - L(t, x^*, u)\}, \quad \text{a.e.} \tag{4.15}$$

Finally, we define $p(\cdot):[0, 1] \to \mathbb{R}^n$:

$$p(t) := (\tilde{p}_1(t), \tilde{p}_2(t))P^{-1}.$$

This function has the required regularity properties, and (2.1) and (2.2) follow from (4.14) and (4.15).

5. Bounded controls. While we are prepared to countenance unbounded controls in optimal control theory, any practical controller design procedure must take account of control magnitude constraints. Such constraints are inevitable; they may describe performance limits on the control actuators, or serve to confine the control and state variables to regions where the given dynamic equations provide an adequate description of the system response.

The presence of constraints can add very significantly to the difficulty of solving an optimal control problem. It is common practice then to replace the constraints by a penalty term. The cost in this case takes the form:

$$J(x(\cdot), u(\cdot)) = \int_0^1 [L_1(t, u(t), x(t)) + \varepsilon|u(t)|^r]\, dt,$$

in which L_1 is the cost integrand of the original, constrained, problem and $\varepsilon|u|^r$ is the penalty term. Here $\varepsilon > 0$ and $r \geqq 1$ are parameters whose magnitudes are adjusted to force the minimizing control to assume values in acceptable regions of the control space.

We should like to know then when inclusion of the term $\varepsilon|u|^r$ in the cost gives rise to bounded controls. In quadratic cost optimal control, where $L_1(t, u, x) = x'Qx$ for some symmetric matrix Q ($Q \geqq 0$), inclusion of $\varepsilon|u|^2$ has this effect for any $\varepsilon > 0$. This is shown by direct calculation of the optimal control [2], an approach which is not available for broader classes of problems.

Our object in this section is to derive from the regularity results of § 2 new criteria for minimizing controllers to be bounded, and in particular for the penalization technique just described to work.

THEOREM 5.1. *Let (x^*, u^*) be a minimizing process. Suppose that, in addition to* (H1)-(H3), *the following hypothesis is satisfied: there exists an integrable function $\gamma(\cdot)$ such that*

$$\max\{|a| + |b|: a \in \partial_x L(t, x^*(t), u^*(t)),\ b \in \partial_u L(t, x^*(t), u^*(t))\} \leqq \gamma(t), \quad \text{a.e.} \tag{5.1}$$

Then u^* is essentially bounded on $[a, b]$.

Proof. We deal only with the case in which (A, B) is controllable. The noncontrollable case is handled by the obvious state reduction techniques (c.f. the proof of Theorem 2.3). Denote by Ω the regular points of u^*. Take a point $\bar{t} \in (a, b) \cap \Omega$ and define $t_{\max} := \sup\{t > \bar{t}: u^* \text{ is essentially bounded on } [\bar{t}, t]\}$. We show presently that u^* is essentially bounded on $[\bar{t}, t_{\max}]$. It will follow from Theorem 2.1 that $t_{\max} \in \Omega$. In view of the definition of $t_{\max}$, this implies that $t_{\max} = b$. Thus u^* is essentially bounded on $[\bar{t}, b]$. Likewise we show that u^* is essentially bounded on $[a, \bar{t}]$, and hence on $[a, b]$.

To show that u^* is essentially bounded on $[\bar{t}, t_{\max}]$, we take a sequence of points $\{t_i\}$ in $(\bar{t}, t_{\max})$ with $t_i \uparrow t_{\max}$. For each i we know that u^* is essentially bounded on $[\bar{t}, t_i]$. Regarding the restriction of (x^*, u^*) to $[\bar{t}, t_i]$ as a minimizer with respect to this time interval and appropriate endpoint constraints, we are justified then in applying the Maximum Principle [3, Thm. 5.1.2]. Accordingly there exists an absolutely continuous function $p_i : [0, 1] \to \mathbb{R}^n$ such that

$$-\dot{p}_i(t) \in p_i(t) \cdot A - \partial_x L, \quad \text{a.e. } [\bar{t}, t_i] \tag{5.2}$$

and

$$-p_i(t) \cdot B \in \partial_u L, \quad \text{a.e. } [\bar{t}, t_i]. \tag{5.3}$$

(In these two inclusions the generalized gradients are evaluated at $(t, x, u) = (t, x^*(t), u^*(t))$.)

Noting that, under the hypotheses, selectors of $t \to \partial_x L(t, x^*(t), u^*(t))$ and $t \to \partial_u L(t, x^*(t), u^*(t))$ are uniformly integrably bounded, we deduce from these inclusions that the numbers $|p_i(1)|$, $i = 1, 2, \cdots$, are uniformly bounded. A "controllability Grammian" argument is involved, precisely along the lines of the passage in the proof of Theorem 2.3 which follows label (4.6).

Once again, using the integrable boundedness of $\partial_x L$, we now obtain from (5.2) and Gronwall's inequality a constant c such that

$$\|p_i\|_{\infty, \bar{t}, t_i} < c \quad \text{for } i = 1, 2, \cdots.$$

It follows now from (5.3), the convexity of $L(t, x^*(t), \cdot)$ and hypothesis (H3) that there exist positive constants c_1 and c_2 such that

$$-\theta(|u^*(t)|) + c_1|u^*(t)| > -c_2, \quad \text{a.e.} \tag{5.4}$$

Let $r_0 > 0$ be such that

$$\theta(r)/r > c_1 + c_2/r_0, \quad \text{whenever } r > r_0.$$

Consider the set of points t on which $|u^*(t)| > r_0$. Then

$$\theta(u^*(t))/|u^*(t)| > c_1 + c_2/|u^*(t)|.$$

By (5.4), such points form a nullset. This establishes that $\|u^*\|_{\infty, a, b} \leqq r_0$, and completes the proof.

It is a simple step now to derive conditions for the penalty term to suppress unboundedness of optimal controls.

COROLLARY 5.2. *Let (x^*, u^*) be a minimizing process for problem (P). Assume that hypotheses* (H1)-(H3) *are satisfied and L has the form*

$$L(t, x, u) = L_1(t, x, u) + \varepsilon|u|^r,$$

in which $L_1 : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a given function, and $r (\geqq 1)$ and $\varepsilon (> 0)$ are given numbers. Suppose further that

- $L_1(t, x, v) \geqq 0$ *for all* $(t, x, u) \in [a, b] \times \mathbb{R}^n \times \mathbb{R}^m$

and

- *given a compact set* $K \subset [a, b] \times \mathbb{R}^n$, *there exists a number* c, *such that*

$$\max \{|a|+|b|: a \in \partial_x L(t, x, u),\ b \in \partial_u L(t, x, u)\} \leqq c_1(1+|u|^r),$$

for all $(t, x) \in K$ *and* $u \in \mathbb{R}^m$. *Then* u^* *is essentially bounded on* $[a, b]$.

Proof. Under the extra hypotheses, the left side of (5.1) is bounded by the integrable function $c_1(1+|u^*(t)|^r)$, where c_1 corresponds to K, the radius of a ball containing graph x^*. Now apply Theorem 5.1.

The corollary says that u^* will be essentially bounded provided the polynomial growth of L and its derivatives, with respect to the control variable, matches that of the penalty term.

Note that in problem $P(k, \varepsilon)$ of § 1, the exponent $r=2$ in the penalty term $\varepsilon|u|^r$ is insufficient; there is an unbounded minimizing control. If however we were to set $r=14$, or more, the theory of this section tells us that minimizing controllers would then be bounded.

Appendix. This appendix provides some results on the decomposition of the dynamic and endpoint constraints

$$\dot{x}(t) = Ax(t) + Bu(t), \quad \text{a.e. } [a, b],$$
$$x(a) = \xi_0, \qquad x(b) = \xi_1,$$

of problem (P). Notice that there is no exogenous term $d(t)$ in the dynamical equations. A is $n \times n$ and B is $n \times m$. It is assumed throughout that (A, B) is controllable and B has full column rank.

LEMMA A.1. *There exist matrices* F, M *and* K ($\det F \neq 0$, $\det M \neq 0$) *of dimension* $n \times n$, $m \times m$, *and* $m \times n$, *respectively, and positive integers* $n_1, \cdots, n_m$ *with the following properties*:

$$n_1 \geqq n_2 \geqq \cdots \geqq n_m \quad \textit{and} \quad m = n_1 + n_2 + \cdots + n_m,$$

and defining

$$\tilde{A} = M^{-1}AM - M^{-1}BFK, \qquad \tilde{B} := M^{-1}BF,$$

we have

$$\tilde{A} = \begin{bmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_n \end{bmatrix}, \qquad B = \begin{bmatrix} \tilde{b}_1 & & 0 \\ & \ddots & \\ 0 & & \tilde{b}_m \end{bmatrix},$$

in which J_i *is the* $n_i \times n_i$ *matrix and* $\tilde{b}_i$ *is the* n_i *vector*

$$J_i = \begin{bmatrix} 0 & 1 & & 0 \\ & & \ddots & \\ 0 & & & 1 \\ 0 & & \cdots & 0 \end{bmatrix}, \qquad \tilde{b}_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

for $i = 1, \cdots, m$.

Proof. The nonsingular matrix M and the "controllability indices" $n_1, \cdots, n_m$ are chosen to reduce the dynamical equations to "controllable companion form" [12]. The nonsingular matrix F, which changes the basis on the control space, is chosen to eliminate off-diagonal entries in the block representation of the input matrix. Finally

K is a feedback matrix chosen to place the poles of the reduced dynamic equations at the origin. Details of the steps involved are supplied in [12].

In what follows, F, M, and K remain as in Lemma A.1. Set $\alpha := n_1$.

Proof of Proposition 3.1. We begin by specifying the matrices and vectors in terms of which the mapping $Z : \mathcal{S} \to \tilde{\mathcal{S}}$ is defined. Recall

$$\mathcal{S} := \{\text{processes } (x, u) \text{ for } (P)\colon x(a) = \xi_0\}$$

and

$$\tilde{\mathcal{S}} := \{z \in W^{\alpha,1}([a, b]; \mathbb{R}^m)\colon D_0^{\alpha-1} z(a) = q\}.$$

For $i = 1, \cdots, m$, the α vector q_i is taken to be

$$q_i := \operatorname{col}\{0, [M^{-1}\xi_0]_i\},$$

i.e., to get the q_i's we partition the vector $M^{-1}\xi_0$ into vectors of dimension $n_1, \cdots, n_m$ ($M^{-1}\xi_0 = \operatorname{col}\{[M^{-1}\xi_0]_1, \cdots, [M^{-1}\xi_0]_m\}$) and add leading zeros if necessary to raise the dimension of each vector $[M^{-1}\xi_0]_i$ to α. Now set $q := \operatorname{col}\{q_1, \cdots, q_m\}$. This describes the boundary condition on elements in $\tilde{\mathcal{S}}$.

Given $(x, u) \in \mathcal{S}$, $Z(x, u)$ is taken to be the unique element $z \in W^{\alpha,1}$ satisfying

$$u(t) = F \operatorname{col}\{D^\alpha z_i(t) + k_i \cdot \operatorname{col}\{D_{\alpha-n_j}^{\alpha-1} z_j(t)\}_{j=1}^m\}_{i=1}^m \tag{A.1}$$

and

$$D_0^{\alpha-1} z(a) = q.$$

Here $k_1, \cdots, k_m$ are the rows of K.

Notice that the right side of (A.1) defines a linear transformation of $(D^\alpha z(t), D_0^{\alpha-1} z(t))$ and accordingly can be written

$$u(t) = FD^\alpha z(t) + GD_0^{\alpha-1} z(t) \tag{A.2}$$

for some matrix G, in accordance with the assertions of Proposition 3.1.

The mapping Z is well-defined since, if (x, u) is fixed, (A.2) amounts to a nondegenerate system of α order differential equations for $z = Z(x, u)$,

$$D^\alpha z(t) + F^{-1} G D_0^{\alpha-1} z(t) = F^{-1} u(t),$$

which together with the boundary condition on lower order derivatives, $D_0^{\alpha-1} z(a) = q$, has a unique solution $z \in W^{1,\alpha}$. Z is a bijection in view of (A.1), and since a control function $u(\cdot)$ uniquely determines a state trajectory which satisfies the boundary condition on elements in $\mathcal{S}$, $x(a) = \xi_0$.

Take $z = Z(x, u)$, and an open subinterval $I \subset [a, b]$. It is evident from (A.1) that $u \in L^\infty(I)$ if and only if $D^\alpha z \in L^\infty(I)$ and, for any integer $\beta \geqq 0$, $u \in C^\beta(I)$ if and only if $D^\alpha z \in C^\beta(I)$.

It remains to check that, if $z = Z(x, u)$,

$$x(t) = HD_0^{\alpha-1} z(t),$$

for some matrix H. We shall show specifically

$$x(t) = M \operatorname{col}\{D_{\alpha-n_i}^{\alpha-1} z_i(t)\}_{i=1}^m \tag{A.3}$$

(i.e., H can be taken to be the matrix representation of the right side of (A.3), viewed as a linear transformation of $D_0^{\alpha-1} z_i(t)$). Let $\tilde{x} = M^{-1} x$ and $\tilde{u} = F^{-1} u$. Then it is easy to show that $\tilde{x}$ and $\tilde{u}$ are related according to

$$\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{B}[\tilde{u} - K\tilde{x}],$$

$$\tilde{x}(a) = M^{-1}\xi_0.$$

Now partition the transformed state vector $\tilde{x}$ into vectors of length $n_1, \cdots, n_m$:

$$\tilde{x} = \operatorname{col}\{\tilde{x}_1, \cdots, \tilde{x}_m\}.$$

Noting the special structure of $\tilde{A}$ and $\tilde{B}$ we deduce that $\tilde{x}_i$ is expressible in terms of an element $w_i \in W^{n_i,1}([a, b]; \mathbb{R})$:

$$\tilde{x}_i(t) = D_0^{n_i-1} w_i$$

and

$$D^{n_i} w_i(t) = \{F^{-1}u(t)\}_i + k_i \operatorname{col}\{D_0^{n_j-1} w_j(t)\}_{j=1}^m.$$

Let $z_1, \cdots, z_m$ be defined to satisfy the differential equations

$$D^{\alpha-n_i} z_i(t) = w_i(t),$$

with boundary conditions

$$D_0^{\alpha-n_i-1} z_i(a) = 0,$$

and write $z = \operatorname{col}\{z_1, \cdots, z_m\}$. Then

$$\tilde{x}_i(t) = D_{\alpha-n_i}^{\alpha-1} z_i(t), \qquad D_0^{\alpha-1} z_i(a) = \operatorname{col}\{0, \tilde{x}_i(a)\}$$

and

$$\tilde{u}(t) = \operatorname{col}\{D^{\alpha} z_i(t) + k_i \cdot \operatorname{col}\{D_{\alpha-n_j}^{\alpha-1} z_j(t)\}_{j=1}^m\}_{i=1}^m.$$

Since $\tilde{x}(t) = M^{-1}x(t)$ and $\tilde{u}(t) = F^{-1}u(t)$, we deduce $z = Z(x, u)$ and (A.3) is true.

REFERENCES

[1] J. BALL AND V. MIZEL, *One-dimensional variational problems whose minimizers do not satisfy the Euler-Lagrange equation*, Arch. Rational Mech. Anal., 90 (1985), pp. 325-388.

[2] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[3] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983; reprinted by CRM, Université de Montréal, 1989.

[4] F. H. CLARKE AND P. D. LOEWEN, *An intermediate existence theory in the calculus of variations*, Annali Scuola Normale Superiole Pisa, to appear.

[5] F. H. CLARKE AND R. B. VINTER, *On the conditions under which the Euler equation or maximum principle hold*, Appl. Math. Optim., 12 (1984), pp. 73-79.

[6] ———, *Regularity properties of solutions to the basic problem in the calculus of variations*, Trans. Amer. Math. Soc., 289 (1985), pp. 73-98.

[7] ———, *Existence and regularity in the small in the calculus of variations*, J. Differential Equations, 59 (1985), pp. 336-354.

[8] ———, *Regularity of solutions to variational problems with polynomial lagrangians*, Bull. Polish Acad. Sci., 34 (1986), pp. 73-81.

[9] ———, *A regularity theory for variational problems with higher order derivatives*, Trans. Amer. Math. Soc., to appear.

[10] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[11] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[12] R. E. KALMAN, *Kronecker invariants and feedback*, in Proc. Conference on Ordinary Differential Equations (Mathematics Research Center, Madison, WI, 1971), Naval Research Laboratory, Washington, DC, 1971.

[13] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1968.

[14] L. TONELLI, *Fondamenti di Calcolo delle Variazioni*, Vols. 1, 2, Zanichelli, Bologna, 1921, 1923.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 999-1048, September 1990

Invited Expository Article

This paper is another in the continuing series of expository papers that were invited by the editors. These papers undergo the same refereeing procedure as do research papers submitted directly by the authors, although the refereeing guidelines are modified to suit the largely expository nature of the paper. Due to the rapid recent technical development of a number of areas in control and optimization, many of the seminal papers are quite specialized and are readily accessible to a limited group of experts only. Moreover, the original motivations and practical importance of the ideas are sometimes difficult to find in the mathematical development. The purpose of these papers is to bring the ideas, techniques, and applications of a few selected areas to the attention of a wider audience, so that their basic importance can be more easily and widely appreciated.

NUMERICAL METHODS FOR STOCHASTIC CONTROL PROBLEMS IN CONTINUOUS TIME*

HAROLD J. KUSHNER†

Abstract. A powerful and usable class of methods for numerically approximating the solutions to optimal stochastic control problems for diffusion, reflected diffusion, or jump-diffusion models is discussed. The basic idea involves a consistent approximation of the model by a Markov chain, and then solving an appropriate optimization problem for the Markov chain model. A general method for obtaining a useful approximation is given. All the standard classes of cost functions can be handled. Here, for illustrative purposes, discounted and average cost per unit time problems with both reflecting and nonreflecting diffusions are concentrated on. Both the drift and the variance can be controlled. Owing to its increasing importance and to lack of material on numerical methods, an application to the control of queueing and production systems in heavy traffic is developed in detail. The methods of proof of convergence are relatively simple, using only some basic ideas in the theory of weak convergence of a sequence of probability measures or random processes. For the deterministic problem, one form of the method reduces to the method of finite elements, but the probabilistic approach allows a much simpler proof of convergence than that usually used for the deterministic problem.

Key words. numerical methods for stochastic control, optimal stochastic control, diffusion approximations, reflected diffusions, weak convergence, martingale measures, Markov chain approximations, ergodic control

AMS(MOS) subject classifications. 93E25, 93E20

1. Introduction. This paper deals with a family of powerful and intuitively appealing numerical methods for a wide variety of stochastic (and even deterministic) control problems. The underlying stochastic processes can be of the diffusion, jump-diffusion, or reflected diffusion types discussed in [20] and [25] or of the sort of reflected diffusion that arises in the heavy traffic modeling and control of queueing and production systems [22], [31], [44]. The basic idea is the "Markov chain approximation" method used in [29] and [35]. The essential idea involves an approximation of the controlled stochastic process by a suitable controlled Markov chain, for which the optimal cost functions are readily computed, and then showing that we can approximate the optimal cost for the original problem arbitrarily closely by the optimal cost for such a controlled chain.

We survey, update, and substantially extend the basic ideas in [29] in a way that allows the coverage of a much wider variety of problems in a somewhat more straightforward way, including relaxed controls, heavy traffic modeling, ergodic control, and

* Received by the editors July 1, 1988; accepted for publication (in revised form) November 20, 1989. This research was supported partially by Air Force Office of Scientific Research contract AFOSR-85-0315, National Science Foundation grant ECS-8505674, and Army Research Office grant DAAL03-86-K-0171.

† Division of Applied Mathematics, Lefschetz Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912.

variance control as well as all the cases dealt with in [29]. The method can handle control over a finite interval, control until a target set is reached, discounted control, optimal stopping and impulsive control, average cost per unit time over an infinite time interval, and various combinations and extensions of these. There are versions that are suitable for (convergent) approximations to nonlinear filtering problems [29, Chap. 7.5], [30], [12], and these can be extended to cover smoothing and interpolation problems. An extension of the filtering approximation covering point processes also appears in [11] and an interesting application of that to reliability theory is in [9]. Unlike classical methods for solving optimal stochastic control problems, smoothness of the solutions to the Bellman-Hamilton-Jacobi (partial differential—or partial differential integral) equations (BHJ) is not needed. On the contrary, the numerical methods to be presented yield numerical approximations to the solutions of such (weak sense) equations by exploiting the fact that their solutions have representations as functionals of diffusions or jump diffusions.

The basic idea of the technique is discussed in § 2. The concept of relaxed controls is very useful in dealing with existence and convergence questions in the calculus of variations and control theory [2], [50]. This technique simplifies the analysis, since (typically) any sequence of relaxed controls has a convergent subsequence (in an appropriate topology; see § 3). The use of this technique for stochastic problems was introduced in [17] and then used in [18] and [37], and elsewhere. In § 3, we define the admissible relaxed stochastic controls. Section 4 contains some definitions concerning weak convergence, and a useful criterion for tightness, as well as a brief discussion of the general method of showing that the limits of our approximating interpolated chains are indeed controlled diffusion processes. Methods for obtaining the approximating chain are given in § 5. For practical programming, efficient methods are essential. Two basic "automatic" schemes are discussed; the first getting the transition probabilities for the chain from the coefficients of a suitably chosen finite-difference approximation to the BHJ equation (whether or not the equation has a smooth solution—the use of finite differences is just a "device" to get the chain: the method is not a finite-difference method), the second involves a form of finite-element method (again, this is just a "device"—the methodology or assumptions of finite elements are not used).

Deterministic control problems are often special cases of the stochastic problems dealt with here. It is instructive to see how the general ideas simplify and can be used for deterministic problems, and for purposes of illustration this is dealt with in § 6 for a special two-dimensional case.

Section 7 treats a problem for a controlled (drift only) diffusion process, with a discounted cost—until a target set is hit or a desired domain is exited. This illustrates the general method of proving tightness and convergence of the controlled approximating chains to an optimal diffusion process. Very similar techniques are used for many other problem types.

In § 8, we treat the problem of optimally controlling a diffusion where the control also appears in the "diffusion" term. The Markov chain method also works here. The proofs require the introduction of the concept of martingale measure [14], [51] to deal with the type of "controlled" stochastic integrals needed for the proper representation of the system. In § 9, we treat the numerical method for the type of reflected diffusion that arises in the modeling of queueing and production systems operating in a so-called heavy traffic environment. The ideal limit model here is a reflected diffusion in the positive orthant of some Euclidean space. Two separate approximation problems arise here, since the "limit reflected diffusion" is not the actual physical model. We must

first find the appropriate limit control problem and an approximation procedure for solving that optimization problem. Then we must show that the optimal value functions for the controlled true physical processes (as the traffic intensity goes to one) converges to the optimal value function for the limit controlled reflected diffusion. In §§ 10 and 11 we treat the reflected diffusion and average cost per unit time problem, respectively. For simplicity, we concentrate on the case of diffusion processes. There are obvious extensions of the method and of all the results to controlled jump diffusions, and to impulsively controlled problems. The singular control problem has some special twists and will be dealt with in a separate paper.

2. Approximation by Markov chains: Introduction. To see how the Markov chain approximation method works, consider our canonical system (2.1) with cost given by (2.2) and where m or $m(\cdot)$ denotes a control policy. The control value u is in U, a compact set, and G^0 is defined in (A2.2) below:

$$dx = b(x, u)dt + \sigma(x)dw, \qquad x \in R^r \quad \text{Euclidean } r\text{-space}, \tag{2.1}$$

$$V(x, m) = E_x^m \int_0^\tau e^{-\beta t} k(x(t), u(t))\, dt + E_x^m\, e^{-\beta\tau} g(x(\tau)), \qquad \beta > 0, \tag{2.2}$$

$$\tau = \inf\{t: x(t) \notin G^0\} \quad \text{if } x(t) \notin G^0, \text{ for some } t < \infty,$$

$$= \infty \quad \text{otherwise.}$$

The control terminology will be made more precise in § 3. E_x^m denotes the expectation under $x(0) = x$ and control policy $m(\cdot)$ used. Assumptions (A2.1) and (A2.2) will be used. Also, $w(\cdot)$ is a standard vector-valued Wiener process, and the dimensions of x, b, σ, and w are compatible.

(A2.1) $b(\cdot, \cdot)$ is continuous and $b(\cdot, u)$, $\sigma(\cdot)$ are Lipschitz continuous, uniformly in u. Write $a(\cdot) = \sigma(\cdot)\sigma'(\cdot)$.

Note that our definition of $a(\cdot)$ is twice the value used in [29].

(A2.2) $k(\cdot)$ and $g(\cdot)$ are continuous and bounded and G is a compact set, which is the closure of its interior G^0, and has a piecewise smooth (differentiable) boundary.

Our purpose in this section is only illustrative. More details and conditions appear in the later sections. Our conditions are also usually stronger than necessary for the convergence of the numerical methods: e.g., usually a weak sense uniqueness condition can replace the Lipschitz condition. $u(\cdot) = m$ is said to be admissible if it is U-valued, measurable, and nonanticipative with respect to $w(\cdot)$. An extended definition of admissibility appears in § 3. The discounted cost (2.1) is chosen here for illustrative purposes and is used in most of the sequel. All the standard forms of the cost function can readily be treated. See [29]. The set G might or might not be part of the original problem statement. In order to numerically solve a control problem it is usually necessary that the process be confined to some bounded set, and some bounding set G might be used for numerical purposes only. Define

$$V(x) = \inf_{m \text{ admissible}} V(x, m). \tag{2.3}$$

The Markov chain approximation. For each $h > 0$, let $\{\xi_n^h, n \geqq 0\}$ be a controlled Markov chain with transition probabilities $p^h(x, y|u)$, $u \in U$, and a discrete state space $R_h^r \in R^r$. Let u_n^h denote the control used at step n, and suppose that an (continuous in x, u) interpolation interval $\Delta t^h(x, u)$ is given. Define $\Delta t_n^h = \Delta t^h(\xi_n^h, u_n^h)$, and set $G_h = R_h^r \cap G^0$, the effective state space of $\{\xi_n^h\}$ until escape from G^0.

Define $\delta\xi_n^h = \xi_{n+1}^h - \xi_n^h$. Let E_n^h denote the expectation given $\{\xi_i^h, u_i^h, i \leqq n\}$. Suppose that for some $\alpha > 0$ and $\xi_n^h = x$,

$$
\begin{aligned}
&E_n^h \delta\xi_n^h = \Delta t^h(x,u) b(x,u) + O(h^\alpha \Delta t^h(x,u)),\\
&E_n^h[\delta\xi_n^h - E_n^h\delta\xi_n^h][\delta\xi_n^h - E_n^h\delta\xi_n^h]' = a(x)\Delta t^h(x,u) + O(h^\alpha \Delta t^h(x,u)),\\
&|\xi_{n+1}^h - \xi_n^h| = O(h).
\end{aligned} \tag{2.4}
$$

We use the $O(h)$ bound in (2.4) for convenience in unifying the presentation. It would not usually hold for cases where the control is of the impulsive, singularly controlled, or hard reflection type. But in all of these cases there are obvious modifications. Approximations for the impulsively controlled problem appear in [29], and the singularly controlled problem [26], [27] would use a combination of the ideas in this paper and those in [38]. Owing to its special nature and the length of this paper, the singular control problem will be dealt with in a separate paper.

Define the interpolated chain $\xi^h(\cdot)$ and control $u^h(\cdot)$ by

$$
\xi^h(t) \equiv \xi_n^h \text{ and } u^h(t) = u_n^h \text{ on } [t_n^h, t_{n+1}^h),\ t_n^h = \sum_0^{n-1} \Delta t_i^h. \tag{2.5}
$$

DEFINITION (to be extended in § 3). A control policy m^h for $\{\xi_n^h\}$ is said to be admissible if the associated u_n^h are U-valued random variables and, with $\{u_n^h\}$ used, the process is still a Markov chain, i.e.,

$$
P\{\xi_{n+1}^h = y | \xi_i^h, u_i^h, i \leqq n\} = p^h(\xi_n^h, y | u_n^h). \tag{2.6}
$$

Let $E_x^{m^h}$ denote the expectation under initial condition x and policy m^h.

Define the discounted cost

$$
V^h(x, m^h) = E_x^{m^h} \sum_0^{N_h - 1} e^{-\beta t_n^h} k(\xi_n^h, u_n^h) \Delta t_n^h + E_x^{m^h} e^{-\beta t_{N_h}^h} g(\xi_{N_h}^h), \tag{2.7}
$$

where $N_h = \min\{n: \xi_n^h \notin G_h\}$ (or $N_h = \infty$ if it is not otherwise defined). Set $\tau_h = t_{N_h}^h$, the escape time of $\xi^h(\cdot)$ from G^0. Then (2.7) can be written

$$
V^h(x, m^h) = E_x^{m^h} \int_0^{\tau_h} e^{-\beta s} k(\xi^h(s), u^h(s))\, ds + E_x^{m^h} e^{-\beta\tau_h} g(\xi^h(\tau_h)). \tag{2.8}
$$

Define

$$
V^h(x) = \inf_{m^h \text{ adm.}} V^h(x, m^h). \tag{2.9}
$$

We use "deterministic" intervals $\Delta t^h(x,u)$ (see below for further remarks). We could work with continuous parameter Markov chains $\{\tilde{\xi}^h(\cdot)\}$ with "infinitesimal" transition probabilities given by (2.6), but where the holding times are exponentially distributed, and with mean duration $\Delta t^h(x,u)$. The results would be exactly the same. For our purposes here there is little advantage in doing so. See [35, § 8], where such models are used for approximations to a jump-diffusion process.

Conditions (2.4) are "local consistency" conditions between $x(\cdot)$ and $\xi^h(\cdot)$. It will turn out that this is the essential quality we require of the chain, irrespective of how it is obtained (or appropriate analogues for the jump-diffusion or "singular" cases). This consistency and the similarity between (2.2) and (2.8) suggest that the optimal values $V^h(x)$ and $V(x)$ will also be close for small h, and this will turn out to be the case. We try to choose $\{p^h(x, y|u)\}$ so that the process $\xi^h(\cdot)$ approximates $x(\cdot)$ "reasonably well," while keeping the computation of $V^h(x)$ tractable.

Computing $V^h(x)$. The dynamic programming equation for (2.7) is

$$V^h(x) = \min_{u \in U} \left[\sum_y e^{-\beta \Delta t^h(x,u)} p^h(x, y|u) V^h(y) + \Delta t^h(x, u) k(x, u) \right], \qquad x \in G_h, \tag{2.10}$$

$$= g(x), \qquad x \notin G_h.$$

There are many methods for computing the $V^h(\cdot)$, although we do not deal with them here. References [32], [42], and [45] concern Gauss–Seidel or successive overrelaxation or other methods that have been found to work well on two- or three-dimensional problems. A promising "aggregation" method is discussed in [3]. The Markov chain approximation scheme has been incorporated into the expert system of Quadrat et al. [6], [7]. There are forms of the multigrid method being investigated for computing $V(\cdot)$ [1]. Although these are not "Markov chain" methods, they appear to have considerable promise, at least for nondegenerate models (2.1).

Convergence of $\{V^h(\cdot)\}$. Under quite general conditions, any subsequence of $\{\xi^h(\cdot)\}$ has a further subsequence that converges to some controlled diffusion with an admissible policy $\tilde{m}$ in the sense of weak convergence (see § 4), and $V^h(x)$ will converge to the cost $V(x, \tilde{m})$ for that diffusion. The limit policy might be a "relaxed" control, even if the $u^h(\cdot)$ are not (see § 3 for the definition). Clearly, $V(x, \tilde{m}) \geqq V(x)$, since $\tilde{m}$ is no better than the optimal control. Then $\underline{\lim}_h V^h(x) \geqq V(x)$. In order to get $V^h(x) \to V(x)$, we need a reverse inequality. To get that, given any $\delta > 0$, we find a special δ-optimal control m^δ for $x(\cdot)$ that has an "adaptation" $m^{h,\delta}$ for use on $\{\xi_n^h\}$ so that $V^h(x, m^{h,\delta}) \to V(x, m^\delta)$. These estimates and $V^h(x) \leqq V^h(x, m^{h,\delta})$ (due to the optimality of $V^h(x)$) yield the desired result. In fact, the convergence will be uniform in $x \in G^0$ under broad conditions.

On $\Delta t^h(x, u)$. For many applications, we might wish to use $\Delta t^h(x, u) = \Delta$, a constant, for example, in applications to the nonlinear filtering problem [29, Chap. 7.5]. As seen in § 5, a suitable chain with constant interpolation intervals can always be found. Fixed $\Delta t^h(x, u) = \Delta$ intervals are also an alternative (but not the only one) when control is over a fixed finite time interval, or whenever the underlying PDE is of the parabolic type [29, Chap. 7]. Actually, a fixed interval can be used with any problem—but it is not usually (numerically) efficient to do so. We comment on this further in § 5 when specific chains are dealt with.

For problems where the control is potentially over an unbounded interval, it is usually both intuitively natural and numerically desirable to let $\Delta t^h(x, u)$ depend on x and u. The usual methods of choosing the chain explicitly yield the $\Delta t^h(x, u)$, so it is obtained automatically. Intuitively, as $|b(x, u)|$ or $|a(x)|$ increases, we want a shorter $\Delta t^h(x, u)$, because the "faster" the dynamics, the shorter the desired "holding times."

3. Relaxed controls. In this section, we define the actual class of admissible controls that will be used. The definitions given here greatly simplify the convergence analysis (in comparison with the methods in [29]). Let $(\Omega, P, \mathcal{F})$ be a probability space and $\mathcal{F}_t$ a sequence of nondecreasing sub-σ-algebras of $\mathcal{F}$. Let $w(\cdot)$ be an $\mathcal{F}_t$-standard vector-valued Wiener process. (That is, $w(t)$ is $\mathcal{F}_t$-measurable and $w(t+s) - w(t)$, $s \geqq 0$, is independent of all $\mathcal{F}_t$-measurable random variables.)

Deterministic controls. A U-valued measurable function $u(\cdot)$ is said to be an *ordinary control* for (3.1):

$$\dot{x} = b(x, u). \tag{3.1}$$

Let $m(\cdot)$ be a measure on the Borel sets of $U \times [0, \infty)$ such that

$$m(U \times [0, t]) = t \quad \text{for all } t \geqq 0. \tag{3.2}$$

For notational convenience, we often write $m(B \times [0, t]) \equiv m(B, t)$. Under (3.2), there is a measure $m_t(\cdot)$ on the Borel sets $\mathcal{U}$ of U such that for Borel B, $m_\cdot(B)$ is Borel measurable, and $m(dc\, dt) = m_t(dc)\, dt$. We say that $m(\cdot)$ is an *admissible relaxed control* for (3.1) and rewrite (3.1) as

$$\dot{x} = \int b(x, c) m_t(dc). \tag{3.3}$$

Any ordinary control $u(\cdot)$ has a representation as a relaxed control where $m_t(dc) = \delta_{u(t)}(c)\, dc$. There is a so-called "chattering lemma" [2], which states that for any relaxed $m(\cdot)$ and associated $x(\cdot)$ there is a sequence of ordinary admissible controls $u^\delta(\cdot)$, each taking only a finite number of values, and associated solutions $x^\delta(\cdot)$ to (3.1), such that $x^\delta(\cdot) \to x(\cdot)$, uniformly on each interval $[0, T]$, and $m^\delta(\cdot) \to m(\cdot)$ (weak topology), where $m^\delta(\cdot)$ is the relaxed control that is equivalent to $u^\delta(\cdot)$.

The advantages provided by the relaxed controls stem from the fact that the dynamics in (3.3) are now linear in the control, and it is much easier to work with questions of approximation, limits of sequences of controls, and existence of optimal controls [2], [17], [50], and similarly for the stochastic case below.

Stochastic controls. Write $\mathcal{F}_t \times \mathcal{B}_t$ for the minimal σ-algebra containing the product sets, where $\mathcal{B}_t$ is the Borel algebra on $[0, t]$, and completed with respect to the null sets ($P \times$ Lebesgue measure) of $\mathcal{F}_\infty \times \mathcal{B}_\infty$. If a U-valued function $u(\cdot)$, when restricted to $\Omega \times [0, t]$, is $\mathcal{F}_t \times \mathcal{B}_t$ measurable, we say that it is an *ordinary admissible control* for the SDE

$$dx = b(x, u)\, dt + \sigma(x)\, dw. \tag{3.4}$$

Such functions $u(\cdot)$ are also called progressively measurable.

Let $m(\cdot)$ be a measure-valued random variable (on the Borel sets of $U \times [0, \infty)$) such that for all t and Borel B

$$m(U, t) = t, \tag{3.5}$$

$$m(B, t) \text{ is } \mathcal{F}_t \text{ measurable.} \tag{3.6}$$

Conditions (3.5) and (3.6) imply the existence of a measure-valued $\mathcal{F}_t$-adapted derivative $m_t(\cdot)$ such that $m(dc\, dt) = m_t(dc)\, dt$. This follows from the following facts. $m(B, \cdot)$ is nondecreasing and Lipschitz continuous (Lipschitz constant less than or equal to 1). Hence for Borel B, $[m(B, t) - m(B, t - \delta)]/\delta$ converges (for almost all t, ω) to an F_t-adapted process $m_t(B)$. Similarly to what we do in the construction of a regular conditional probability distribution, a version of $m_\cdot(B)$ can be chosen such that the convergence is for all Borel B (for almost all ω, t), and $m_t(\cdot)$ is a measure on the Borel sets of U.

We call $m(\cdot)$ an *admissible relaxed control* for (3.3), and for such controls, the model (3.7) replaces (3.3):

$$dx = dt \int b(x, c) m_t(dc) + \sigma(x)\, dw. \tag{3.7}$$

If $u(t) = u_0(x(t), t)$ for some Borel function $u_0(\cdot)$, we call $u(\cdot)$ an *admissible feedback control.* The only important fact concerning the filtration $\mathcal{F}_t$ with respect to which admissibility is defined is that $w(\cdot)$ is an $\mathcal{F}_t$-Wiener process. If $\mathcal{F}_t$ is not important,

we often say that $(m(\cdot), w(\cdot))$ is an *admissible pair* or that $m(\cdot)$ *is admissible with respect to* $w(\cdot)$.

The following result is well known [17], [18], and the proof is a natural analogue of the proof of the chattering lemma for the deterministic case.

THEOREM 3.1. *Assume* (A2.1), *and let* $(m(\cdot), w(\cdot))$ *be an admissible pair. Then* (3.7) *has a unique* (*strong sense*) *solution. Also the probability law of* $(x(\cdot), w(\cdot))$ *is determined by that of* $(m(\cdot), w(\cdot))$. *For any* $T<\infty$, $\delta>0$, *there is an ordinary admissible control* $u^\delta(\cdot)$, *which is piecewise constant and takes only a finite number of values and is such that if* $x^\delta(\cdot)$ *and* $x(\cdot)$, *respectively, correspond to* $u^\delta(\cdot)$, *and* $m(\cdot)$, *respectively, then for any bounded continuous real-valued* $f(\cdot)$

$$
\begin{aligned}
&P\left\{\sup_{t\leqq T}|x^\delta(t)-x(t)|>\delta\right\}\overset{\delta}{\to}0,\\
&E\left|\int_0^T f(s,u^\delta(s))\,ds-\int_0^T\int f(s,c)m_s(dc)\,ds\right|\to 0.
\end{aligned}
\tag{3.8}
$$

For $m(\cdot)$ an admissible relaxed control, the $V(x, m)$ of (2.2) is written as

$$
V(x,m)=E_x^m\int_0^\tau\int e^{-\beta t}k(x(t),c)m_t(dc)\,dt+E_x^m\,e^{-\beta\tau}g(x(\tau)).
\tag{3.9}
$$

The reader is reminded that the cost depends on the joint distribution of $(m(\cdot), w(\cdot))$ and not only on $m(\cdot)$, but we omit the $w(\cdot)$ from the notation for simplicity.

Controls for $\{\xi_n^h\}$. The U-valued sequence $\{u_n^h\}$ is said to be an *ordinary admissible control* if (2.6) holds. If there are Borel functions $u_n(\cdot)$ such that $u_n^h=u_n(\xi_n^h)$, then it is said to be an *ordinary admissible feedback control.* A sequence $m_n^h(\cdot)$ of measure-valued (on the Borel sets of U) random variables is an *admissible relaxed control* if $m_n^h(U)\equiv 1$ and

$$
P\{\xi_{n+1}^h=y|\xi_i^h, m_i^h, i\leqq n\}=\int p^h(\xi_n^h, y|c)m_n^h(dc).
$$

Throughout the paper, we will use only ordinary admissible controls for the $\{\xi_n^h\}$. Nevertheless, it is often important to use the relaxed control terminology and the relaxed control representation of the ordinary control, since the limits of the "relaxed control representations" of the ordinary controls might not be ordinary controls, but only relaxed controls.

Define $m^h(\cdot)$ by its derivative $m_t^h(\cdot)$: $m_t^h(\cdot)=m_n^h(\cdot)$ on $[t_n^h, t_{n+1}^h)$, and write $m^h(dc\,dt)=m_t^h(dc)\,dt$. We rewrite (2.8) in terms of relaxed controls as

$$
V^h(x,m^h)=E_x^{m^h}\int_0^{\tau_h}\int e^{-\beta s}k(\xi^h(s),c)m_s^h(dc)\,ds+E_x^{m^h}\,e^{-\beta\tau}g(\xi^h(\tau_h)).
\tag{3.10}
$$

For any ordinary control $\{u_n^h\}$, we define the relaxed control equivalent by $m_n^h(dc)=\delta_{u_n^h}(c)\,dc$.

4. Weak convergence. In the first part of this section, we discuss the concepts and results in the theory of weak convergence, which will be useful to us in the sequel. In the second part, these results are applied to characterize the limits of the interpolated processes $\{\xi^h(\cdot), m^h(\cdot)\}$, as $h\to 0$. Let $\{X_n\}$ and X be random variables with values in a metric space S, and with induced probabilities $\{P_n\}$ and P on the σ-algebra $\mathcal{S}$ of S. If $Ef(X_n)\to Ef(X)$ for every bounded real-valued continuous function $f(\cdot)$, then

we say that $\{X_n\}$ (or $\{P_n\}$) *converges weakly* to X (or P), and write $X_n \Rightarrow X$ (or $P_n \Rightarrow P$). An important extension (used in §§ 6–9 below) [4, Thm. 5.1] is Theorem 4.1.

THEOREM 4.1. *Let $X_n \Rightarrow X$ and suppose that $f(\cdot)$ is bounded, real-valued, and measurable with discontinuity set D_f. If $P\{D_f\} = 0$ (where P is the measure induced by X), then $Ef(X_n) \to Ef(X)$ and $f(X_n) \Rightarrow f(X)$.*

The following theorems and definitions provide the basis of the subject of weak convergence [4], [15], [28]. The sequence $\{X_n\}$ (or $\{P_n\}$) is said to be *tight* (relatively compact) if for each $\varepsilon > 0$, there is a compact set $K_\varepsilon \subset S$ such that

$$\inf_n P\{X_n \in K_\varepsilon\} \geqq 1 - \varepsilon.$$

If the X_n are vector-valued, then tightness is equivalent to $\lim_N \sup_n P\{|X_n| \geqq N\} = 0$. One part of Prohorovs' theorem [4, Thm. 6.1] states the following.

THEOREM 4.2. *Let $\{X_n\}$ be tight in $(S, \mathscr{S})$. Then for each subsequence, there is a further subsequence $\{X_{n_i'}\}$ and an X such that $X_{n_i'} \Rightarrow X$.*

For analytical purposes, it is often more convenient to work with probability one convergence than with weak convergence. If only the distributions of the random variables or functions are of interest, then the underlying probability space is unimportant and can be chosen so that the following theorem of Skorokhod holds [25, Thm. 2.7, Chap. 1].

THEOREM 4.3 (Skorokhod representation). *Let S be a complete and separable metric space with metric $d(\cdot, \cdot)$ and let $X_n \Rightarrow X$ on S. Then there is a probability space $(\tilde{\Omega}, \tilde{P}, \tilde{\mathscr{F}})$ with S-valued random variables $\tilde{X}_n$, $\tilde{X}$, defined on it such that for all Borel sets A in S*

$$\tilde{P}\{\tilde{X}_n \in A\} = P\{X_n \in A\},\ \tilde{P}\{\tilde{X} \in A\} = P\{X \in A\}$$

and $d(\tilde{X}_n, \tilde{X}) \to 0$ with probability 1.

Let $\mathscr{M}(T)$, $T < \infty$, denote the set of measures $m(\cdot)$ on the Borel sets of $U \times [0, T]$ that satisfy $m(U \times [0, t]) \equiv m(U, t) = t$, for all $t \leqq T$, and with weak topology. On $\mathscr{M}(\infty)$ we use the strongest topology which coincides with that of each $\mathscr{M}(T)$ on $[0, T]$. Since each $\mathscr{M}(T)$, $T < \infty$, is compact, so is $\mathscr{M}(\infty)$, and any sequence of $\mathscr{M}(\infty)$-valued random variables has a weakly convergent subsequence. Such a "compact weak" topology will be used for all the measure-valued functions. The topologies are metrizable.

Let $D^k[0, \infty)$ denote the space of R^k-valued functions on $[0, \infty)$ that are right continuous and have left-hand limits and all endowed with the Skorokhod topology [4, § 1.4], [28, Chap. 2].

The Skorokhod topology is metrizable so that $D^k[0, \infty)$ is a complete and separable metric space under that metric. *Very loosely speaking*, a compact set A in $D^k[0, \infty)$ is characterized by the facts that on any bounded interval $[0, T]$, the elements are uniformly bounded, and for any $\varepsilon > 0$ the number of discontinuities larger than ε are bounded and the functions are equicontinuous between the "discontinuities."

The processes and random variables of interest in this paper are mostly $\{\xi^h(\cdot)\}$ with paths in $D^r[0, \infty)$, $\{m^h(\cdot)\}$ with values in $\mathscr{M}(\infty)$, and stopping times $\{\tau_h\}$ with values in the compact space $[0, \infty] = \bar{R}$. For processes $\{Y^n(\cdot)\}$ with paths in $D^k[0, \infty)$, a very convenient criterion for tightness (due to Aldous and Kurtz [28, Thm. 2.7]) is given by the following theorem. Let $\mathscr{F}_t^n$ denote the minimal σ-algebra generated by $\{Y^n(s), s \leqq t\}$.

THEOREM 4.4. *Suppose that for each $T < \infty$,*

$$\lim_{N \to \infty} \sup_n P\left\{\sup_{t \leqq T} |Y^n(t)| \geqq N\right\} = 0. \tag{4.1}$$

For each $T<\infty$, let

$$\lim_{\delta\to 0}\overline{\lim_{n}}\sup_{\tau\leqq T} E\min\{1,|Y^n(\tau+\delta)-Y^n(\tau)|^\beta\}=0, \tag{4.2}$$

for some $\beta>0$, where τ ranges over all $\mathcal{F}_t^n$-stopping times. Then $\{Y^n(\cdot)\}$ is tight in $D^k[0,\infty)$. Under (4.2), (4.1) *is implied by*

$$\lim_{N\to\infty}\sup_{n} P\{|Y^n(t)|\geqq N\}=0 \text{ for each } t. \tag{4.3}$$

If for all t, $Y^n(t)$ takes values in a complete and separable metric space S_1 with metric $d(\cdot)$, then (4.3) *and* (4.2) *are replaced by: for each t and $\varepsilon>0$ there is a compact $K_{\varepsilon,t}\subset S_1$ such that*

$$\inf_{n} P\{Y^n(t)\in K_{\varepsilon,t}\}\geqq 1-\varepsilon, \tag{4.3'}$$

$$\lim_{\delta\to 0}\overline{\lim_{n}}\sup_{\tau\leqq T} E\min\{1,d^\beta(Y^n(\tau+\delta),Y^n(\tau))\}=0 \quad \text{for some } \beta>0. \tag{4.2'}$$

Tightness of $\{\xi^h(\cdot), m^h(\cdot)\}$. Let E_n^h and E_t^h denote the conditional expectation with respect to the σ-algebras generated by $\{\xi_i^h, m_i^h, i\leqq n\}$ and $\{\xi^h(s), m^h(s), s\leqq t\}$, respectively.

THEOREM 4.5. *Assume* (A2.1) *and* (2.4), *and let $m^h(\cdot)$ be the admissible relaxed control representation of a sequence $\{u_n^h\}$ of ordinary admissible controls for $\{\xi_n^h\}$. Then $\{\xi^h(\cdot), m^h(\cdot)\}$ is tight in $D^r[0,\infty)\times\mathcal{M}(\infty)$, and the limits of $\{\xi^h(\cdot)\}$ are continuous processes.*

Proof. For simplicity of development, we let $\sup_{x,c}\Delta t^h(x,c)\xrightarrow{h}0$. The general case is handled in the same way. We first show (4.3) for $Y^n(\cdot)$ replaced by $\xi^h(\cdot)$. We have (mod $O(h)$), with $\xi^h(0)=x$ and for some constant K

$$\begin{aligned} E|\xi^h(t)-x|^2 &= E\left|\sum_{t_n^h\leqq t}(\delta\xi_n^h-E_n^h\delta\xi_n^h+E_n^h\delta\xi_n^h)\right|^2 \\ &\leqq KE\left[\sum_{t_n^h\leqq t}(\Delta t_n^h)\right]^2+KE\sum_{t_n^h\leqq t}\Delta t_n^h, \end{aligned} \tag{4.4}$$

which yields (4.3).

Relation (4.2) also follows from (2.4) and the boundedness of $b(\cdot)$ and $a(\cdot)$, and a calculation similar to (4.4). The $\{m^h(\cdot)\}$ is always tight in the compact set $\mathcal{M}(\infty)$. The fact that the paths of the limits of $\{\xi^h(\cdot)\}$ are continuous follows from: (a) the piecewise linear interpolations of $\xi^h(\cdot)$ are continuous and differ from $\xi^h(\cdot)$ by $O(h)$, and (hence) are also tight in $D^r[0,\infty)$; (b) a sequence of processes in $D^r[0,\infty)$ with continuous paths that converge weakly can only converge to a process with continuous paths, by the properties of the Skorokhod topology. □

Weak convergence of $\{\xi^h(\cdot), m^h(\cdot), \tau_h\}$. By the tightness proved in Theorem 4.5, each subsequence of $\{\xi^h(\cdot)\}$ has itself a subsequence that converges weakly. The next theorem shows that the limits are actually controlled diffusion processes. This fact plays an important role in the proofs that $V^h(x)\to V(x)$. Let $\phi_j(\cdot)$ be bounded and continuous functions on $U\times[0,\infty)$, and define $(m,\phi)_t=\int_0^t\int\phi(s,c)m(dc\,ds)$. Define $\mathcal{L}^c$, the differential generator of (2.1) with $u=c$: for continuous real-valued $f(\cdot)$ having compact support and continuous second partial derivatives,

$$\mathcal{L}^c f(x)=f_x'(x)b(x,c)+\frac{1}{2}\sum_{i,j}f_{x_ix_j}(x)a_{ij}(x).$$

The technique used in Theorem 4.6 is widely used to characterize the limits of weakly convergent sequences as solutions to an appropriate martingale problem.

THEOREM 4.6. *Assume* (2.4) *and* (A2.1) *and let* τ_h *be* $\mathcal{F}_t^h$*-stopping times. Let* $m^h(\cdot)$ *be the admissible relaxed control representation of the interpolated* (*intervals* Δt_n^h) *sequence of admissible ordinary controls* $\{u_n^h\}$. *Suppose that* $\{\xi^h(\cdot), m^h(\cdot), \tau_h\}$ *converges weakly to* $(x(\cdot), m(\cdot), \tau)$. *Then there is a filtration* $\mathcal{F}_t$ *and an* $\mathcal{F}_t$*-standard Wiener process* $w(\cdot)$ *such that* τ *is an* $\mathcal{F}_t$*-stopping time,* $m(\cdot)$ *is admissible* (*i.e.*, $(m(\cdot), w(\cdot))$ *is an admissible pair*), *and*

$$dx = \int b(x, \alpha) m_t(d\alpha)\, dt + \sigma(x)\, dw. \tag{4.5}$$

Proof. Let p, q, t_i, $i \leqq q$, t, and s be arbitrary, but with $t_i \leqq t \leqq t+s$ and such that $P\{\tau = t\} = 0$. We will show that for arbitrary smooth real-valued functions $f(\cdot)$ and $h(\cdot)$ with compact support,

$$\begin{aligned} &Eh(x(t_i), \tau I_{\{\tau \leqq t\}}, (\phi_j, m)_{t_i}, i \leqq q, j \leqq p) \\ &\quad \cdot \left[f(x(t+s)) - f(x(t)) - \int_t^{t+s} \int \mathcal{L}^c f(x(v)) m(dc\, dv) \right] = 0. \end{aligned} \tag{4.6}$$

For the moment, suppose that (4.6) holds. Let $\mathcal{F}_t$ denote the minimal σ-algebra, which measures $\{x(s), m_s(\cdot), \tau I_{\{\tau \leqq s\}}, s \leqq t\}$. Then the arbitrariness of $h(\cdot)$, t_i, $\phi_j(\cdot)$, and (4.6) imply that

$$E_{\mathcal{F}_t}\left[f(x(t+s)) - f(x(t)) - \int_t^{t+s} \int \mathcal{L}^c f(x(v)) m(dc\, dv) \right] = 0.$$

Hence

$$f(x(t)) - \int_0^t \int \mathcal{L}^c f(x(s)) m_s(dc)\, ds \equiv M_f(t)$$

is an $\mathcal{F}_t$-martingale for each $f(\cdot)$ of the chosen type. But this implies that there is an $\mathcal{F}_t$-Wiener process[1] $w(\cdot)$ such that (4.5) holds [25, p. 73], [49]. Thus we need only establish (4.6).

We now use the Skorokhod representation (Theorem 4.3) so that all weak convergences become with probability one (w.p.1) convergences in the topologies of $D^r[0, \infty)$, $\mathcal{M}(\infty)$ or $\bar{R}$, as appropriate. Note that if $y_n(\cdot) \to y(\cdot)$ in the Skorokhod topology, and $y(\cdot)$ is continuous, then the convergence is uniform on bounded time intervals. We now prove (4.6) in the scalar case only, for notational convenience. By (2.4)

$$E_n^h f(\xi_{n+1}^h) - f(\xi_n^h) = \int f_x(\xi_n^h) b(\xi_n^h, c) m_n^h(dc) \Delta t_n^h + \frac{1}{2} f_{xx}(\xi_n^h) \sigma^2(\xi_n^h) \Delta t_n^h + O(h^\alpha \Delta t_n^h).$$

This yields

$$\begin{aligned} E_t^h f(\xi^h(t+s)) - f(\xi^h(t)) &= E_t^h \int_t^{t+s} \int f_x(\xi^h(v)) b(\xi^h(v), c) m^h(dc\, dv) \\ &\quad + \frac{1}{2} E_t^h \int_t^{t+s} f_{xx}(\xi^h(v)) \sigma^2(\xi^h(v))\, dv + \delta^h(t, t+s) \\ &= E_t^h \int_t^{t+s} \mathcal{L}^c f(\xi^h(v)) m^h(dc\, dv) + \delta^h(t, t+s), \end{aligned} \tag{4.7}$$

[1] If $a(x)$ is not uniformly positive definite, then we might have to augment the probability space by adding an independent Wiener process.

where δ^h is bounded and goes to zero as $h \to 0$. Thus

$$Eh(\xi^h(t_i), \tau_h I_{\{\tau_h \leqq t\}}, (\phi_j, m^h)_{t_i}, i \leqq q, j \leqq p)$$
$$\cdot \left[f(\xi^h(t+s)) - f(\xi^h(t)) - \int_t^{t+s} \int \mathscr{L}^c f(\xi^h(v)) m^h(dc\, dv) \right] \xrightarrow{h} 0. \tag{4.8}$$

Finally, (4.6) follows from (4.8) and the weak convergence. □

A representation of $\{\xi_n^h\}$ with martingale driving terms. We now obtain a representation for $\{\xi_n^h\}$ that will be very useful in obtaining the limit results in §§ 7–11. Letting the control sequence be $\{u_n^h\}$ and defining $\beta_n^h = (\xi_{n+1}^h - \xi_n^h) - E_n^h(\xi_{n+1}^h - \xi_n^h)$, we have

$$\xi_{n+1}^h = \xi_n^h + b(\xi_n^h, u_n^h)\Delta t_n^h + \beta_n^h + O(h^\alpha \Delta t_n^h), \tag{4.9}$$

where $\operatorname{cov}_n^h \beta_n^h = a(\xi_n^h)\Delta t_n^h + O(h^\alpha \Delta t_n^h)$, by (2.4).

We now represent $\{\beta_n^h\}$ in terms of "white noise." To understand the scheme, first suppose that $\sigma(x)$ has a uniformly bounded inverse $\sigma^{-1}(x)$, and define $\delta W_n^h = \sigma^{-1}(\xi_n^h)\beta_n^h$. Then ($\operatorname{cov}_n^h$ denotes the conditional covariance, analogous to E_n^h)

$$\begin{aligned} &\operatorname{cov}_n^h \delta W_n^h = I\Delta t_n^h + O(h^\alpha \Delta t_n^h), \qquad |\delta W_n^h| \xrightarrow{h} 0, \\ &E_n^h \beta_n^h (\delta W_n^h)' = \sigma(\xi_n^h)\Delta t_n^h + O(h^\alpha \Delta t_n^h). \end{aligned} \tag{4.10}$$

Define $W^h(\cdot)$ by

$$W^h(t) = \sum_{t_{n+1}^h \leqq t} \delta W_{n-1}^h, \qquad W_n^h = \sum_0^{n-1} \delta W_i^h.$$

We can now write (4.9) in terms of the "white noise" sequence $\{\delta W_n^h\}$. The importance will be seen in Theorem 4.7 and in §§ 7–11. Now drop the invertibility assumption.

For notational simplicity we will only treat the case where $\sigma(\cdot)$ is a square $(r \times r)$ matrix, and not the general case. We follow the scheme in Chapter 6.6 of [6]. Write $a(\xi_n^h) = P_n^h (D_n^h)^2 P_n^{h\prime}$, where D_n^h is diagonal $\{d_{n1}^h, \cdots, d_{nr}^h\}$ and P_n^h is an orthogonal matrix, both random. For the $\alpha \in (0, 1)$, define

$$I_n^h = \operatorname{diag}\{d_{n1}^h I_{\{d_{n1}^h > h^\alpha\}}, \cdots\}.$$

(Note that our $a(\cdot)$ is twice the $a(\cdot)$ used in [6].) Let $\psi(\cdot)$ denote an R^r-valued standard Wiener process independent of $\{\xi_n^h, u_n^h\}$, and set $\delta\psi_n^h = \psi(t_{n+1}^h) - \psi(t_n^h)$. Extend the definition of E_n^h and cov_n^h so that they include conditioning on $\delta\psi_i^h$, $i < n$. Define (D^{-1} denotes the pseudo-inverse)

$$\delta W_n^h = (D_n^h)^{-1} I_n^h (P_n^h)' \beta_n^h + (I - I_n^h)\delta\psi_n^h. \tag{4.11}$$

Then we can write

$$\begin{aligned} &\operatorname{cov}_n^h \delta W_n^h = I\Delta t_n^h + O(h^\alpha)\Delta t_n^h, \\ &E_n^h \beta_n^h (\delta W_n^h)' = \sigma(\xi_n^h)\Delta t_n^h + O(h^\alpha \Delta t_n^h), \\ &\beta_n^h = \sigma(\xi_n^h)\delta W_n^h + \varepsilon_n^h, \end{aligned} \tag{4.12}$$

where the continuous parameter interpolation $\varepsilon^h(\cdot)$ of $\{\varepsilon_n^h\}$ will converge weakly to the "zero" process. In fact $E_n^h \varepsilon_n^h = 0$ and $\operatorname{cov}_n^h \varepsilon_n^h = O(h^\alpha)\Delta t_n^h$, and $\sup_{n,\omega} \varepsilon_n^h \to 0$ as $h \to 0$. We write (4.9) as

$$\xi_{n+1}^h = \xi_n^h + b(\xi_n^h, u_n^h)\Delta t_n^h + \sigma(\xi_n^h)\delta W_n^h + \varepsilon_n^h + O(h^\alpha \Delta t_n^h). \tag{4.13}$$

Let $m^h(\cdot)$ denote the relaxed control representation of the continuous parameter interpolation of $\{u_n^h\}$, with interpolation intervals $\{\Delta t_n^h\}$.

We can state the following theorem.

THEOREM 4.7. *Assume* (2.4) *and* (A2.1). *Then* $\{\xi^h(\cdot), W^h(\cdot), \varepsilon^h(\cdot), m^h(\cdot)\}$ *is tight in* $D^{3r}[0,\infty)\times\mathcal{M}(\infty)$. *If the limit of some convergent subsequence is denoted by* $(x(\cdot), w(\cdot), \varepsilon(\cdot), m(\cdot))$, *then* $\varepsilon(\cdot)\equiv 0$ *and the rest satisfy* (3.7), *where* $m(\cdot)$ *is an admissible relaxed control.* (*The filtration* F_t *is that determined by the limit processes.*)

Remark on the proof. The proof is very similar to that of Theorem 4.6. $\{W^h(\cdot)\}$ is tight in $D^r[0,\infty)$. The fact that $\varepsilon^h(\cdot)\Rightarrow$zero process follows from its local properties. Then, in the arguments of $h(\cdot)$ in Theorem 4.6, replace $\xi^h(u)$ by $(\xi^h(u), W^h(u))$ whenever it appears. Let $f(\cdot)$ depend on ξ and w and note that (vector case notation)

$$E_n^h f(\xi_{n+1}^h, W_{n+1}^h) - f(\xi_n^h, W_n^h) = f_x'(\xi_n^h, W_n^h) b(\xi_n^h, u_n^h)\Delta t_n^h + O(h^\alpha \Delta t_n^h)$$
$$+\frac{1}{2}\Delta t_n^h\left[\sum_{i,j} f_{x_i x_j}(\xi_n^h, W_n^h) a_{ij}(\xi_n^h) + \sum_i f_{w_i w_i}(\xi_n^h, W_n^h) + \sum_{i,j} f_{x_i w_j}(\xi_n^h, W_n^h)\sigma_{ij}(\xi_n^h)\right].$$

Then substitute into the vector form of (4.7) and (4.8) and take weak limits. On taking limits, $\mathcal{L}^c$ will be replaced by the operator of the pair $(x(\cdot), w(\cdot))$ satisfying (3.7) under $m(\cdot)$.

An extension. There is an extension of Theorem 4.7 that will be needed in §§ 8 and 9, where $a(x)$ has the form

$$a(x) = \sum_i^q \sigma_i(x)\sigma_i'(x). \tag{4.14}$$

In the subsequent development we suppose that each σ_i is square $(r\times r)$ (but this restriction can be dropped) and Lipschitz continuous. The associated system is

$$dx = \int b(x, c) m_t(dc)\, dt + \sum_{i=1}^q \sigma_i(x)\, dw^i(s), \tag{4.15}$$

where $\{w^1(\cdot), \cdots, w^q(\cdot)\}$ are mutually independent standard R^r-valued Wiener processes. We will construct approximations $\{W^{1,h}(\cdot), \cdots, W^{q,h}(\cdot)\}$.

We follow a procedure that is similar to the technique used in § 5.2 below to construct an approximating Markov chain. Consider the $q+1$ systems

$$\dot{x}^0 = \int b(x^0, c) m_t(dc), \quad dx^i = \sigma_i(x^i) dw^i, \quad 0 < i \leqq q. \tag{4.16}$$

The Markov chain $\{\xi_n^h\}$ will be constructed on the discrete state space used in § 2. Suppose that $\hat{p}_i^h(x, y \mid c)$, $i = 0, \cdots, q$, is the transition function for the Markov chain approximation associated with the ith system in (4.16), with associated interpolation times $\Delta t_i^h(x, c)$, and let the consistency conditions analogous to (2.4) hold. As will be seen in § 5.2, the interpolation times take the form $h^2/Q_{ih}(x, c) = \Delta t_i^h(x, c)$, where Q_{ih} is a weighted sum of the coefficients b_i and a_{ij}, and we assume this form here. Even though the systems for $i \geqq 1$ do not depend on c, we retain the original notation for consistency.

Define $Q_h(x, c) = \sum_{i=0}^q Q_{ih}(x, c)$, and define

$$p^h(x, y \mid c) = \sum_{i=0}^q \hat{p}_i^h(x, y \mid c) Q_{ih}(x, c)/Q_h(x, c).$$

Then $\Delta t^h(x, c) = h^2/Q_h(x, c)$. In fact, the transition probability p^h for the Markov chain approximation to (4.15) can always be decomposed into the above form.

Now, let $\xi_n^h = x$, $u_n^h = u$, and represent the random transition from x to ξ_{n+1}^h as follows. Choose system i, $0 \leqq i \leqq q$ (i.e., the transition probability $\hat{p}_i^h(x, y \mid u)$), with conditional (given the past data) probability $Q_{ih}(x, u)/Q_h(x, u)$, and then use the chosen $\hat{p}_i^h(x, y \mid u)$ to get ξ_{n+1}^h. Let $I_n^{h,i}$ denote the indicator function that system i is selected. Let $E_n^{h,i}$ denote the conditional expectation given $\{\xi_k^h, u_k^h, k \leqq n\}$ and the fact that system i is selected at time n, and extend the definition of E_n^h so that it includes conditioning on the systems selected at times $i \leqq n$. We can write

$$\xi_{n+1}^h = \xi_n^h + \Delta t_n^h b(\xi_n^h, u_n^h) + \sum_{i=1}^{q} \beta_n^{h,i} + O(h^\alpha \Delta t_n^h) + \bar{\varepsilon}_n^h, \tag{4.17}$$

where

$$\beta_n^{h,i} = [(\xi_{n+1}^h - \xi_n^h) - E_n^{h,i}(\xi_{n+1}^h - \xi_n^h)]I_n^{h,i},$$

$$\bar{\varepsilon}_n^h = \sum_{i=0}^{q} [E_n^{h,i}(\xi_{n+1}^h - \xi_n^h) \cdot I_n^{h,i} - E_n^h(\xi_{n+1}^h - \xi_n^h)]I_n^{h,i}.$$

The $\{\beta_n^{h,i}, i = 1, \cdots, q, n < \infty\}$ are orthogonal, and

$$E_n^h \beta_n^{h,i}(\beta_n^{h,j})' = \delta_{ij}\sigma_i(\xi_n^h)\sigma_i'(\xi_n^h)\Delta t_n^h + O(h^\alpha \Delta t_n^h).$$

Also, the error process defined by $\sum_{t_n^h \leqq t} \bar{\varepsilon}_n^h$ converges weakly to the zero process as $h \to 0$. By introducing q independent processes $\psi^i(\cdot)$, $i \leqq q$, similar to what was done in Theorem 4.7 above, we can define $\delta W_n^{h,i}$ such that

$$\beta_n^{h,i} = \sigma_i(\xi_n^h)\delta W_n^{h,i} + \varepsilon_n^{h,i}, \tag{4.18}$$

and where the process defined by $\sum_{t_n^h \leqq t} \varepsilon_n^{h,i}$ converges weakly to the zero process.

We have the following extension of Theorem 4.7.

Define

$$W^{h,i}(t) = \sum_{t_n^h \leqq t} \delta W_{n-1}^{h,i}.$$

Assume $b(\cdot, \cdot)$ is continuous, $b(\cdot, u)$, $\sigma_i(\cdot)$ is Lipschitz continuous uniformly in $u \in U$, and let (2.4) hold for the $\hat{p}_i^h$ for each subsystem $i = 0, \cdots, q$. Then $\{\xi^h(\cdot), m^h(\cdot), W^{h,i}(\cdot), i \leqq q\}$ is tight. Any weak limit $(x(\cdot), m(\cdot), w^i(\cdot), i \leqq q)$ satisfies (4.15). The $\{w^i(\cdot)\}$ are mutually independent standard $\mathcal{F}_t$-Wiener processes and $m(\cdot)$ is admissible, where $\mathcal{F}_t = \sigma\{x(s), w(s), m(s), s \leqq t\}$.

5. Methods for the construction of an approximating Markov chain. Any method of constructing the approximating Markov chain for which (2.4) holds can be used. We discuss two methods here. The first uses a "finite-difference" technique, the second includes the first as well as a "finite-element" method. The first method has the advantage of being essentially automatic—an important consideration in any program. In § 6, we specialize the second approach to a deterministic problem, in which form the general motivation for our choices becomes easy to see. The construction of the chain is guided by two principles: satisfaction of (2.4) and ease of solution of the optimization problem for the chain. The construction of the chain can readily be modified to handle impulsive or singular control problems or cases where there are "hard" reflections. In §§ 9 and 10 modifications for cases with boundary reflections are developed.

5.1. A finite-difference method. Let R_h^r denote the h-grid on R^r defined by $R_h^r = \{x: x = \sum_1^r n_i e_i h, \ n_i \text{ integers}\}$, where e_i denotes the unit vector in the ith coordinate direction. By carefully choosing a finite-difference approximation for $\mathcal{L}^c f(x)$ for

arbitrary $f(\cdot)$, we obtain directly the transition probabilities from the coefficients of the $f(x+n_ie_ih+n_je_jh)$ in the finite-difference expansion, where $n_i=1$, 0, or -1. The idea is to use a finite-difference method that reflects the actual direction (or probability distribution of the direction) of motion of the dynamical system.

Following the method of [29, pp. 91 ff.], we use

$$\begin{aligned} &f_{x_i}(x)\to[f(x+e_ih)-f(x)]/h \quad \text{if } b_i(x,c)\geqq 0,\\ &f_{x_i}(x)\to[f(x)-f(x-e_ih)]/h \quad \text{if } b_i(x,c)<0, \end{aligned} \tag{5.1}$$

$$f_{x_ix_i}(x)\to[f(x+e_ih)+f(x-e_ih)-2f(x)]/h^2. \tag{5.2}$$

For $i\neq j$ and $a_{ij}(x)\geqq 0$, we use

$$\begin{aligned} f_{x_ix_j}(x)\to{}&[2f(x)+f(x+e_ih+e_jh)+f(x-e_ih-e_jh)]/2h^2\\ &-[f(x+e_ih)+f(x-e_ih)+f(x+e_jh)+f(x-e_jh)]/2h^2. \end{aligned} \tag{5.3}$$

For $i\neq j$ and $a_{ij}(x)<0$, we use

$$\begin{aligned} f_{x_ix_j}(x)\to{}&-[2f(x)+f(x+e_ih-e_jh)+f(x-e_ih+e_jh)]/2h^2\\ &+[f(x+e_ih)+f(x-e_ih)+f(x+e_jh)+f(x-e_jh)]/2h^2. \end{aligned} \tag{5.4}$$

Apart from allowing the choices to depend on the signs of the dynamical terms, the finite-difference schemes are standard. The reasons for the choices will be clear below.

Assume that

$$a_{ii}(x)-\sum_{j:\,j\neq i}|a_{ij}(x)|\geqq 0, \quad \text{for all } i, x, \tag{5.5}$$

and define

$$Q_h(x,c)=\sum_i a_{ii}(x)-\sum_{\substack{i\neq j\\ i,j}}\frac{|a_{ij}(x)|}{2}+h\sum_i|b_i(x,c)|.$$

Condition (5.5) will be weakened below. Define the interpolation interval

$$\Delta t^h(x,c)=h^2/Q_h(x,c) \tag{5.6}$$

and the transition probabilities (5.7)

$$\begin{aligned} &p^h(x,x\pm e_ih\,|\,c)=\left[\frac{a_{ii}(x)}{2}-\sum_{j:j\neq i}\frac{|a_{ij}(x)|}{2}+hb_i^{\pm}(x,c)\right]\Big/Q_h(x,c),\\ &p^h(x,x+e_ih+e_jh\,|\,c)=p^h(x,x-e_ih-e_jh\,|\,c)=a_{ij}^{+}(x)/2Q_h(x,c),\\ &p^h(x,x-e_ih+e_jh\,|\,c)=p^h(x,x+e_ih-e_jh\,|\,c)=a_{ij}^{-}(x)/2Q_h(x,c). \end{aligned} \tag{5.7}$$

The $p^h(x,y\,|\,c)$ are zero for all nonlisted values of y.

The $p^h(x,y\,|\,c)$ and $\Delta t^h(x,c)$ were obtained as follows. Substitute the finite-difference expressions (5.1)–(5.4) for the partial derivatives in $\mathscr{L}^cf(x)+k(x,c)=0$, multiply all terms by h^2 (to clear the denominator), and then divide all terms in the resulting expression by the coefficient of $f(x)$ (which is, in fact, $Q_h(x,c)$) to get the expression

$$f(x)=\sum_y p^h(x,y\,|\,c)f(y)+\Delta t^h(x,c)k(x,c). \tag{5.8}$$

Equation (5.8) suggests that the chain with transition probabilities $p^h(x,y\,|\,c)$ and interpolation intervals $\Delta t^h(x,c)$ provides an approximation to $x(\cdot)$. In fact, if $\{\xi_n^h\}$ is the chain with transition probabilities (5.7), and $\Delta t^h(x,c)$ is defined by (5.6) then it is readily verified that (2.4) holds.

On the choice (5.1)–(5.4). The form (5.1) was used to ensure that the $b_i(x, c)$ contribute the correct bias to the mean direction of "flow" of $\{\xi_n^h\}$, and similarly for the other choices.

On (5.5). The condition (5.5) is used to ensure that the $p^h(x, x \pm e_i h \mid c)$ are all $\geqq 0$, but it fails if some of the a_{ij} are large compared to a_{ii}. In that case, there are several alternatives. We can rotate the grid R_h^r so that the coordinate lines are more closely aligned with the eigendirections of the $a(x)$. If these directions change substantially as x varies, curvilinear coordinates can be used. A third method allows the value of h to depend on the coordinate direction. Refer to Fig. 5.1. For illustrative purposes, let $a_{11} = 1$, $a_{12} = 2$, $a_{22} = 5$. Then (5.5) fails. But if we use h_i, $i = 1, 2$, for the difference intervals in the coordinate directions e_1, e_2, respectively, and use the appropriate finite differences (exactly as (5.1)-(5.4), but with $h_i e_i$ replacing he_i), then instead of (5.5), we require that

$$\frac{a_{11}}{h_1^2} - \frac{a_{12}}{h_1 h_2} \geqq 0, \qquad \frac{a_{22}}{h_2^2} - \frac{a_{12}}{h_1 h_2} \geqq 0. \tag{5.9}$$

Any ratio $2 \leqq h_2/h_1 \leqq 2\frac{1}{2}$ will work.

Another alternative is illustrated in Fig. 5.2, and is also a special case of the method of the next subsection. The transitions in (5.7) are only to the nearest neighbors. Greater versatility is obtained if we can also go to the "next to the nearest neighbors."

On simplifying the $\Delta t^h(x, c)$, $p^h(x, y \mid c)$. For computational purposes, the fact that $\Delta t^h(x, c)$ depends on c might be troubling. As seen below, we can often just drop the c-dependence. Assume

$$\inf_x \left[\sum_i a_{ii}(x) - \sum_{\substack{i,j \\ i \neq j}} \frac{|a_{ij}(x)|}{2} \right] > 0 \tag{5.10a}$$

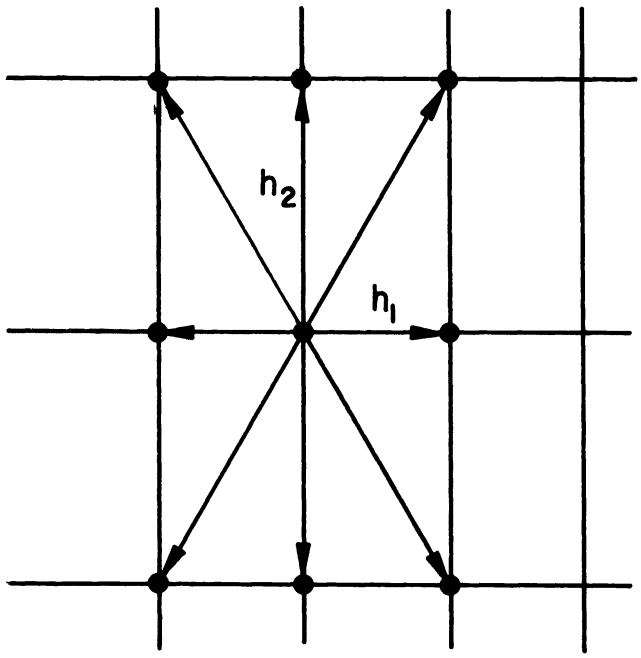


FIG. 5.1. *The connections when ξ_n^h communicates only to nearest neighbors and $h_1 \neq h_2$.*

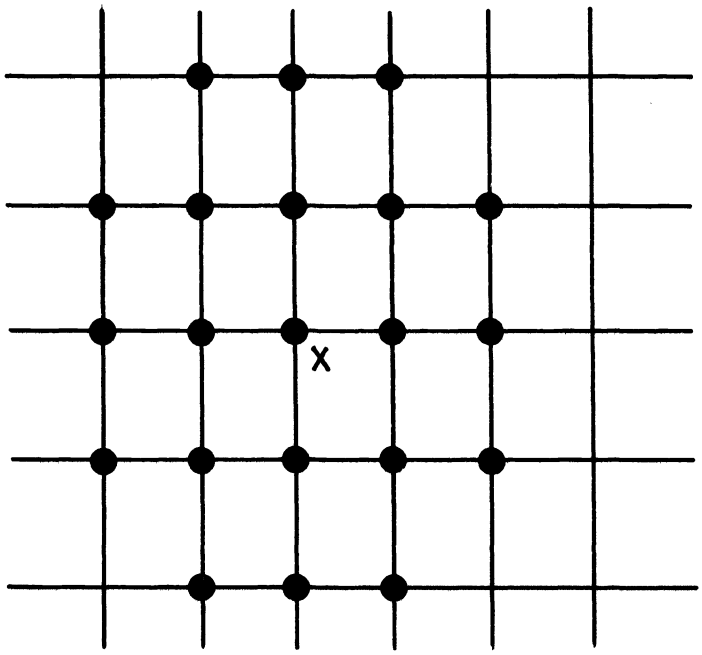


FIG. 5.2. *The points to which x communicates, if next-to-nearest neighbors are included.*

and define

$$\Delta t^h(x) = h^2 \Big/ \left[\sum_j a_{ii}(x) - \sum_{\substack{i,j \\ i \neq j}} \frac{|a_{ij}(x)|}{2}\right]. \tag{5.10b}$$

Then $\Delta t^h(x) - \Delta t^h(x, c) = O(h)\Delta t^h(x)$ and (2.4) still holds with $\Delta t^h(x)$ used in lieu of $\Delta t^h(x, c)$. Then, we can minimize in (2.10) using $\Delta t^h(x)$ and still get the correct limits as $h \to 0$.

Similarly, the dependence of the denominator of the expression for $p^h(x, y \mid c)$ on c might be troubling from a numerical point of view. This dependence can be removed by modifying the transition probabilities as follows. Define

$$\bar{Q}_h(x) = \sup_{c \in U} Q_h(x, c)$$

and define $p^h(x, y \mid c)$ for $y \neq x$, as in (5.7) but with $\bar{Q}_h(x)$ replacing $Q_h(x, c)$, and set $\Delta t^h(x) = h^2/\bar{Q}_h(x)$. With the *new values* of $p^h(x, y \mid c)$ (for $x \neq y$) used, define

$$p^h(x, x \mid c) = 1 - \sum_{y \neq x} p^h(x, y \mid c).$$

The new $\Delta t^h(x)$ and transition probabilities also satisfy (2.4). Under (5.10a) this yields $p^h(x, x \mid c) = O(h)$. The numerical methods for getting $V^h(x)$ in (2.10) usually converge faster as the "mass spreads faster." Then, the larger $p^h(x, x \mid c)$ is, the slower the convergence rate would be. Similar remarks apply to the constant $\Delta t^h(x, c) = \Delta$ interval case discussed next.

Constant interpolation intervals $\Delta t^h(x, c)$. For problems where control is only over a fixed finite interval or where the actual evolution of estimates over time is of interest (such as for nonlinear filtering problems), it is often convenient to use constant

interpolation intervals. Some relevant schemes are discussed in Chapter 7 of [29]. The general idea is very similar to what was done above—or will be done in the next subsection. One quick method is to replace $Q_h(x, c)$ by its maximum (over x and c) and define a new $p^h(x, x \mid c)$ by the difference between unity and the sum over y of the new $p^h(x, y \mid c)$, $y \neq x$. Then $\Delta t^h(x) = h^2/\sup_{x,c} Q_h(x, c)$.

5.2. A finite-element method. We now illustrate a method that covers both the finite-element (see § 6) and finite-difference methods. The scheme can easily be adjusted to accommodate "nonlocal" movements of the chain, as might occur (for example) in singular or impulsive control. For illustrative purposes, we describe it in R^2 only. For a small scalar h, let G_h denote a given set of discrete points in R^2 and for each $x \in G_h$, let $\{v_i(x), i, x \in G_h\}$ be a given collection of vectors. The collection of vectors $\{x + hv_i(x), \text{all } i, x \in G_h\}$ "triangulate" R^2 as in Fig. 5.3, where the "arms" emanating from x are $\{x + hv_i(x)\}$.

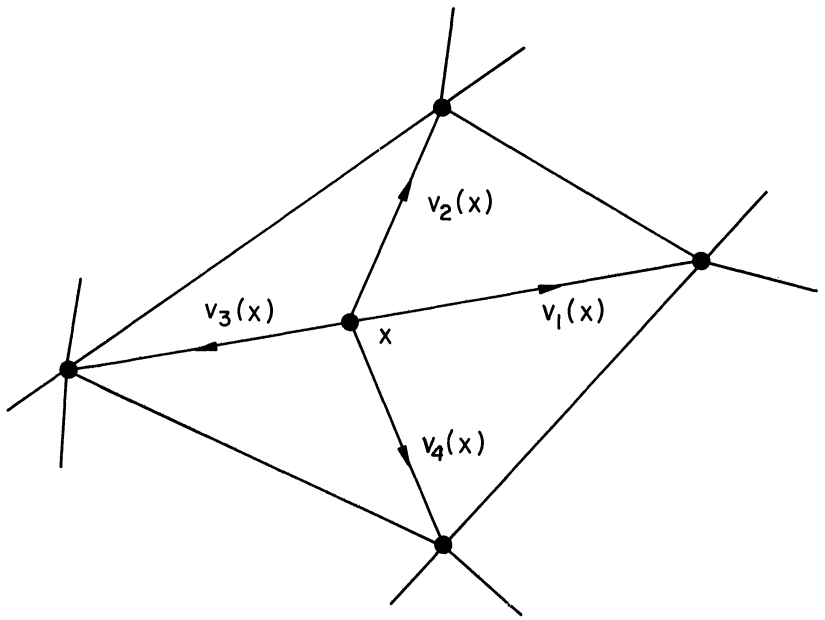


FIG. 5.3. *Triangulation via the* $\{v_i(x)\}$.

Suppose that there are $p_i^h(x, c) \geqq 0$ and $\Delta t^h(x, c) > 0$ such that $\sum_i p_i^h(x, c) = 1$ and for some $\alpha > 0$

$$
\begin{aligned}
&h \sum_i p_i^h(x, c) v_i(x) = b(x, c) \Delta t^h(x, c) + O(h^\alpha \Delta t^h(x, c)), \\
&h^2 \sum_i p_i^h(x, c) v_i(x) v_i'(x) = a(x) \Delta t^h(x, c) + O(h^\alpha \Delta t^h(x, c)).
\end{aligned} \tag{5.11}
$$

The transition probabilities defined by $p^h(x, y \mid c) = p_i^h(x, c)$ for $y = x + hv_i(x)$, $x \in G_h$, and the associated chain $\{\xi_n^h\}$ and interpolation intervals $\Delta t^h(x, c)$ satisfy the consistency condition (2.4).

The $p_i^h(x, c)$ and $\Delta t(x, c)$ can often be conveniently found in the following way. Let the $v_i(x)$ occur in opposite pairs. That is, for each $v_i(x)$, there is $v_j(x) = -v_i(x)$.

If possible, choose $q_i^0(x, c) \geqq 0$, $q_i^1(x) \geqq 0$ such that

$$b(x, c) = \sum_i q_i^0(x, c) v_i(x),$$

$$a(x) = \sum_i q_i^1(x, c) v_i(x) v_i'(x), \tag{5.12}$$

$$q_i^1(x) = q_j^1(x) \quad \text{if } v_i(x) = -v_j(x).$$

Then set

$$p_i^h(x, c) = \frac{[hq_i^0(x, c) + q_i^1(x)]}{\sum_j [hq_j^0(x, c) + q_j^1(x)]} = p^h(x, x + hv_i(x) \mid c),$$

$$\Delta t^h(x, c) = \frac{h^2}{\sum_j [hq_j^0(x, c) + q_j^1(x, c)]}. \tag{5.13}$$

Note that (5.13) reduces to the values obtained in § 5.1 when G_h is the h-grid in R^2 and $\{v_i(x)\} = \{\pm e_i, e_i \pm e_j, -e_i \pm e_j, i, j, i \neq j\}$ and $q_i^0(x, c) = b_i^{\pm}(x, c)$ for $v_i(x) = \pm e_i$, and are zero otherwise, and the $q_i^1(x)$ are chosen in the obvious way.

In § 6, we will see how this greatly simplifies for the deterministic problem, and the relationship to a finite-element method will be discussed.

6. The deterministic discounted problem. In order to illustrate the main ideas of the approximation methods, we now treat a discounted deterministic case where the system and cost are

$$\dot{x} = \int b(x, c) m_t(dc), \tag{6.1}$$

$$V(x, m) = \int_0^\infty c^{-\beta t} k(x(t), c) m_t(dc)\, dt, \tag{6.2}$$

and $b(\cdot)$ and $k(\cdot)$ are bounded and continuous, with $b(\cdot, c)$ Lipschitz continuous in x, uniformly in c. Let $\Delta t^h(x, c)$ be continuous and satisfy $k_2 h \geqq \Delta t^h(x, c)$, $k_2 > 0$, and $\Delta t^h(x, c) \xrightarrow{h} 0$. Refer to Fig. 6.1, where the sides of the triangle are $O(h)$. The dynamic

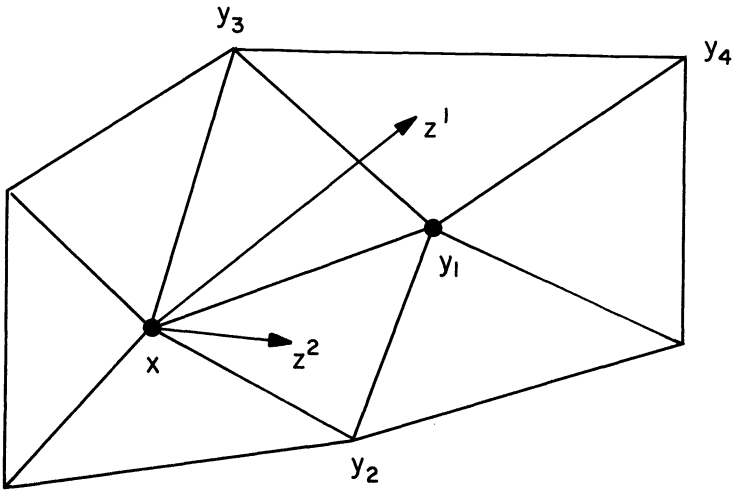


FIG. 6.1. *A finite-element approximation.* $z^2 = b(x, c_2)\Delta t^h(x, c_2) + x = z(x, c_2)$. $z^1 = b(x, c_1)\Delta t^h(x, c_1) + x = z(x, c_1)$.

programming equation for the discrete approximation is (6.3):

$$V^h(x) = \min_{c \in U} [e^{-\beta \Delta t^h(x,c)} V^h(x + b(x,c)\Delta t^h(x,c)) + \Delta t^h(x,c) k(x,c)]. \tag{6.3}$$

Now suppose that we approximate $V^h(\cdot)$ by a piecewise linear function—with the approximation in each triangle of Fig. 6.1 being linear. With use of control value c at node x, we use $z(x,c)$ to denote the canonical point $x + b(x,c)\Delta t^h(x,c)$ reachable from x in time $\Delta t^h(x,c)$, e.g., z^1 or z^2 in Fig. 6.1, where $c = c_1$ or c_2, respectively. Let $Y^h(x,c)$ denote the corners of the triangle in which $z(x,c)$ lies, e.g., $Y^h(x,c_2) = \{x, y_1, y_2\}$ in Fig. 6.1. For $y \in Y^h(x,c)$, let $p^h(x,y \mid c)$ denote the weights that yield $z(x,c)$ (e.g., the weights yielding z^2 in Fig. 6.1 as a convex combination of (x, y_1, y_2)). Clearly $p^h(x,y \mid c) \geqq 0$ and $\sum_{y \in Y^h(x,c)} p^h(x,y \mid c) = 1$. Let $\{\xi_n^h\}$ denote the controlled Markov chain whose transition function is $p^h(x,y \mid c)$, $y \in Y^h(x,c)$. With the "piecewise linear" approximation to $V^h(\cdot)$ used, we can rewrite (6.3) as (for x a vertex of a triangle)

$$V^h(x) = \min_{c \in U} \left[e^{-\beta \Delta t^h(x,c)} \sum_{y \in Y^h(x,c)} p^h(x,y \mid c) V^h(y) + \Delta t^h(x,c) k(x,c) \right]. \tag{6.4}$$

It is clear that (6.4) represents a finite-element approximation.

Calculating the local "statistics" of $\{\xi_n^h\}$, we have (where $E_{n,c}^h$ denotes the expectation given ξ_i^h, u_i^h, $i \leqq n$, and $u_n^h = c$)

$$\begin{aligned} &E_{n,c}^h \xi_{n+1}^h - \xi_n^h = b(\xi_n^h, c)\Delta t^h(x,c) = O(h), \\ &\operatorname{cov}_{n,c}^h (\xi_{n+1}^h - \xi_n^h) = O(h^2). \end{aligned} \tag{6.5}$$

Let $u^h(x)$ minimize in (6.4), and let $u_n^h = u^h(\xi_n^h)$, $\Delta t_n^h = \Delta t^h(\xi_n^h, u_n^h)$. Let $m^h(\cdot)$ denote the relaxed control representation of the continuous parameter interpolation (intervals Δt_n^h) of $\{u_n^h\}$. Then, as in § 4, $\{\xi^h(\cdot), m^h(\cdot)\}$ is tight, and if $(x(\cdot), \tilde{m}(\cdot))$ is the limit of any weakly convergent subsequence, then

$$\dot{x} = \int b(x,c) \tilde{m}_t(dc).$$

Also, if $\xi_0^h = x$, then for that subsequence (indexed by h_n)

$$V^{h_n}(x) \to \int_0^\infty e^{-\beta t} \int k(x(t), c) \tilde{m}_t(dc)\, dt = V(x, \tilde{m}).$$

The probabilistic interpretation is just a device used to study the finite-element approximation for this originally deterministic problem.

Clearly $V(x, \tilde{m}) \geqq V(x) = \inf_{m \text{ adm.}} V(x,m)$.

Next, we wish to show that $V(x, \tilde{m}) = V(x)$. Let $\bar{m}(\cdot)$ denote the optimal admissible (deterministic) relaxed control for (6.1), (6.2). For each $\delta > 0$ and $T_\delta < \infty$, where $T_\delta \to \infty$ as $\delta \to 0$, there is a $\Delta > 0$ and an admissible ordinary (deterministic) control $\bar{u}^\delta(\cdot)$, which is constant on the intervals $[i\Delta, i\Delta + \Delta)$, $i = 0, 1, \cdots$, and is such that for $\bar{x}^\delta(\cdot)$ and $\bar{x}(\cdot)$ corresponding to $\bar{u}^\delta(\cdot)$ and $\bar{m}(\cdot)$, respectively, we have

$$\sup_{t \leqq T_\delta} |\bar{x}^\delta(t) - \bar{x}(t)| < \delta, \qquad (\bar{x}^\delta(\cdot), \bar{m}^\delta(\cdot)) \overset{\delta}{\to} (\bar{x}(\cdot), \bar{m}(\cdot)).$$

We let $\bar{m}^\delta(\cdot)$ denote the relaxed control representation of $\bar{u}^\delta(\cdot)$. Thus, to get $V(x, \tilde{m}) = V(x)$, we need only show that $V(x, \tilde{m}) \leqq V(x, \bar{m}^\delta) = V(x, \bar{u}^\delta)$, for each $\delta > 0$.

We now apply $\bar{u}^\delta(\cdot)$ to $\{\xi_n^h\}$, as follows. Define a control sequence $\{\bar{u}_n^h\}$ for the controlled Markov chain $\{\xi_n^h\}$ in the following way. Let h be small enough so that

$\Delta \gg \inf_{x,c} \Delta t^h(x, c)$. Define a sequence $\{\bar{u}_n^h\}$ recursively by $\bar{u}_n^h = \bar{u}^\delta(0)$ for n such that $t_n^h < \Delta$. Use $\bar{u}_n^h = \bar{u}^\delta(i\Delta)$ for all n such that $t_n^h \in [i\Delta, i\Delta + \Delta)$. Let $\{\bar{\xi}_n^h\}$ denote the associated Markov chain (instead of $\{\xi_n^h\}$), with interpolation $\bar{\xi}^h(\cdot)$, and let $\bar{u}^{\delta,h}(\cdot)$ denote the continuous parameter interpolation (intervals Δt_n^h) of $\{\bar{u}_n^h\}$. Then $\{\bar{\xi}^h(\cdot), \bar{u}^{\delta,h}(\cdot)\} \Rightarrow (\bar{x}^\delta(\cdot), \bar{u}^\delta(\cdot))$. Also $V^h(x, \bar{u}^{\delta,h}) \to V(x, \bar{u}^\delta)$. Since, by optimality of $m^h(\cdot)$, $V^h(x, \bar{u}^{\delta,h}) \geqq V^h(x, m^h) = V^h(x) \to V(x, \tilde{m})$, we conclude that $V^h(x) \to V(x)$, as desired.

A finite-element method for the deterministic problem is discussed in [16] and [21]. The scheme in these papers is essentially that given above, and the probabilistic approach gives a substantially simpler convergence proof. For the stochastic problem, a form of finite-element method is discussed in [43].

7. Convergence of the numerical method for a discounted cost problem. In this section we prove the convergence $V^h(x) \to V(x)$ for system (2.1) or (in the relaxed control form) (3.7) and cost function (2.2), with the approximating chain satisfying (2.4). The stopping or boundary set G is used since, for any practical numerical method, the state space needs to be bounded. The existence of the boundary poses some problems for the convergence and these are discussed below, together with conditions (A7.1), (A7.2), which guarantee the convergence. The discounted problem with stopping on first hitting the boundary of a given set was selected as a canonical problem—with which the basic technique could be readily illustrated.

The dynamic programming equation for $V^h(x)$ is

$$
\begin{aligned}
V^h(x) &= \min_{c \in U} \left[e^{-\beta \Delta t(x,c)} \sum_y p^h(x, y \mid c) V^h(y) + \Delta t^h(x, c) k(x, c) \right], \qquad x \in G_h, \\
&= g(x), \quad x \notin G_h.
\end{aligned} \tag{7.1}
$$

Let $\bar{u}^h(\cdot)$ denote the minimizing control in (7.1) and define $\bar{u}_n^h = \bar{u}^h(\xi_n^h)$. Let m_n^h and $m^h(\cdot)$ denote the relaxed control representations: that is, $m_n^h(dc)$ is the measure concentrated at the point $\bar{u}^h(\xi_n^h)$ and $m_t^h = m_n^h$ on $[t_n^h, t_{n+1}^h)$. Let N_h and τ_h denote the first exit times of $\{\xi_n^h\}$ and $\xi^h(\cdot)$, respectively from G_h.

Discussion of $\tau = \lim \tau_h$. By the results of § 4, the sequence $\{\xi^h(\cdot), m^h(\cdot), W^h(\cdot), \tau_h\}$ is tight. Let $(x(\cdot), m(\cdot), w(\cdot), \tau)$ denote the limit of a weakly convergent subsequence, also indexed by h. Then $(x(\cdot), m(\cdot), w(\cdot))$ satisfy (3.7), where $(m(\cdot), w(\cdot))$ is an admissible pair, and $\tau I_{\{\tau \leqq \cdot\}}$ is nonanticipative with respect to $w(\cdot)$. We always have

$$
E_x^{m^h} \int_0^{\tau_h} \int e^{-\beta t} k(\xi^h(t), c) m_t^h(dc)\, dt \to E_x^m \int_0^\tau \int e^{-\beta t} k(x(t), c) m_t(dc)\, dt, \tag{7.2}
$$

$$
E_x^{m^h} e^{-\beta \tau_h} g(\xi^h(\tau_h)) \to E_x^m e^{-\beta \tau} g(x(\tau)). \tag{7.3}
$$

But it is not always true that τ is the first hitting time of ∂G for the limit process $x(\cdot)$. An example can be seen in the deterministic case illustrated in Fig. 7.1, where $\xi^h(\cdot) \to x(\cdot)$, but τ_h does not go to the first exit time of limit. The problem with the case depicted in Fig. 7.1 is that $x(\cdot)$ is *tangent* to the boundary ∂G at the point of first contact. Such a situation must be avoided, at least w.p.1, if τ is to equal the first exit time. For an arbitrary continuous function $\phi(\cdot)$ with $\phi(0) \in G$, let $\hat{\tau}(\phi(\cdot))$ denote the first time that $\phi(\cdot)$ hits ∂G (set $\hat{\tau}(\phi(\cdot)) = \infty$, if $\phi(t) \notin \partial G$, all $t < \infty$). As illustrated in Fig. 7.1, the function $\tau(\phi(\cdot))$ is not continuous at all $\phi(\cdot)$ (in the topology determined by the sup norm over each interval $[0, T]$). If the function $x(\cdot)$ drawn in Fig. 7.1 were a sample path of a Wiener process or a solution to a (scalar-valued in the example) stochastic differential equation with a nondegenerate covariance, then the law of the

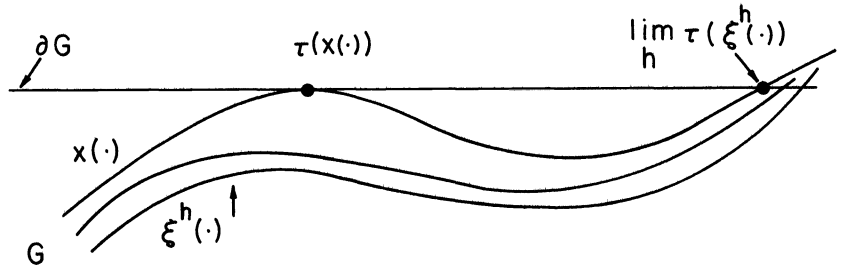


FIG. 7.1. *An example of noncontinuity of* $\tau(\cdot)$. $\xi^h(\cdot) \to x(\cdot)$.

iterated logarithm for such processes implies that if $x(\cdot)$ hits ∂G at τ, then on any set $[\tau, \tau+\delta]$, $\delta > 0$, it crosses ∂G infinitely often w.p.1. This implies that $\tau = \hat{\tau}(x(\cdot))$ w.p.1 and, hence, that

$$V^h(x) = V^h(x, m^h) \to V(x, m). \tag{7.4}$$

Note that $V(x, m) \geqq V(x)$.

We use the following two conditions concerning the boundary ∂G:

(A7.1) The compact set G is the closure of its interior and ∂G is piecewise continuously differentiable.

(A7.2) $\hat{\tau}(\cdot)$ is continuous (in the topology described in the last paragraph) w.p.1 relative to the measures induced by the limit processes $x(\cdot)$, under all admissible controls.

Discussion of (A7.2). First, let $a(x)$ be uniformly positive definite in G. Suppose that there is an open cone C and an $\varepsilon > 0$ such that for each $y \in \partial G$, we have $\{x : x - y \in C, |x - y| < \varepsilon\} \cap G^0 =$ empty set. Then we say that ∂G satisfies the open cone condition, and (via [13, Thm. 13.8]) (A7.2) holds.

Verifying (A7.2) for the degenerate case is more difficult, and we usually check it in each case. Further discussion appears in [29, pp. 64-66]. Define $\tau' = \inf\{t : x(t) \notin \bar{G}\}$, and let S denote the set of points $y \in \partial G$ such that $P_y\{\tau' > 0\} \neq 0$ for some limit process $x(\cdot)$ satisfying (3.7). Frequently, the boundary ∂G can be broken into several pieces, each considered separately; e.g., (a) a section where the orientation and noise is such that the considerations for the nondegenerate problem above work; (b) a section where the dynamics either guarantee (A7.2) or where escape is impossible (due, say, to the sign of the "velocity"), and (c) the remaining section. In many cases, the "remaining section" consists of a finite set of isolated points that are either not accessible or are in S. Consider, for example, the case depicted in Fig. 7.2, where $dx_1 = x_2 dt$, $dx_2 = u\,dt + dw$. The only questionable points are (α) and (β), and these can be shown not to be in S by means of tests such as Theorem 6.1 of [48] (see also [29, p. 66]).

If the set G is not precisely given or can be altered slightly without losing the meaning of the problem, then a slight alteration to the stopping rule yields (A7.2). This is usually the case when G is chosen largely for numerical reasons—to guarantee a bounded state space and a "finite" numerical algorithm. The alternative stopping rule is *randomized stopping.*

Randomized stopping. Recall the definition of S given in the above discussion. Let $q(x) \geqq 0$ denote a continuous function on G^0 that is nonzero only on $N_\varepsilon(S) \cap G^0$, where $N_\varepsilon(S)$ is the ε-neighborhood of S, and that goes to infinity as $x \to \partial S$. Then stop

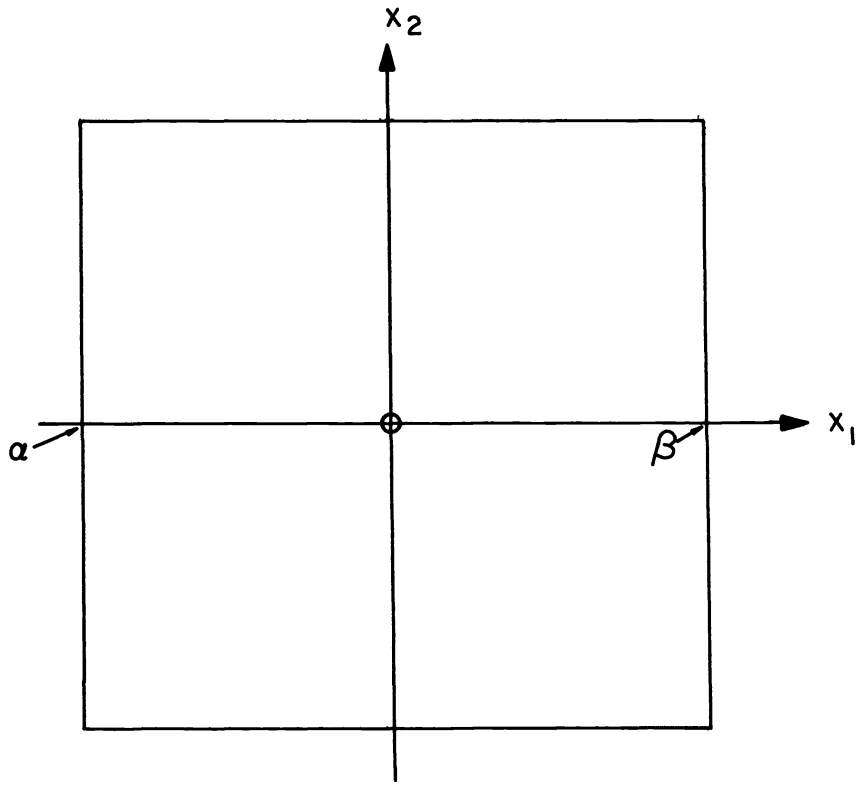


FIG. 7.2. *A degenerate case.* $dx_1 = x_2 dt$, $dx_2 = u\,dt + dw$.

ξ_n^h at time n w.p.1 if $\xi_n^h \notin G^0$ and with probability $1-(\exp-[q(\xi_n^h)\Delta t_n^h])$ if $\xi_n^h \in G^0$. The stopping cost is $g(\xi_n^h)$. The analogous situation for the diffusion (3.7) is to stop either on hitting ∂G or with "stopping rate" $q(x(t))$ at time t, with stopping cost $g(x(t))$. Letting τ_h and τ continue to denote the stopping times, Theorem 7.1 below holds without using (A7.2) under this randomized stopping rule. Although we continue to use (A7.2), the above variations should be kept in mind.

The convergence theorem.

THEOREM 7.1. *Assume* (2.4), (A2.1), (A2.2), (A7.1), (A7.2). *Then* $V^h(x) \to V(x)$.

Proof. In the proof, entities such as $(\xi^h(\cdot), W^h(\cdot), m^h(\cdot), \tau^h)$ or $(x(\cdot), w(\cdot), m(\cdot), \tau)$ that are grouped together are related via the dynamical equations, and the controls will be admissible. Let h index a weakly convergent subsequence with $\{\xi^h(\cdot), W^h(\cdot), m^h(\cdot), \tau_h\} \Rightarrow (x(\cdot), w(\cdot), m(\cdot), \tau)$, where $m^h(\cdot)$ is optimal for $\xi^h(\cdot)$. Then (7.4) holds. Thus we need only prove that for this (or any) weakly convergent subsequence

$$\overline{\lim_h}\, V^h(x) \leqq V(x). \tag{7.5}$$

We will do this by a procedure which is analogous to that used for the deterministic problem at the end of § 6. Let $\bar{m}(\cdot)$ be admissible with respect to $w(\cdot)$ and such that $\bar{x}(\cdot)$ and $\bar{\tau}$ are the associated solution and stopping time and $V(x, \bar{m}) = V(x)$. We need to approximate $\bar{m}(\cdot)$ in such a way that it can be applied to $\{\xi_n^h\}$. First note the following fact. Let $\tilde{m}^\delta(\cdot)$ be admissible with respect to $w(\cdot)$, and let $\tilde{x}^\delta(\cdot)$ and $\tilde{\tau}^\delta$ be the associated solution and stopping time. Then if $(\tilde{m}^\delta(\cdot), w(\cdot)) \Rightarrow (\tilde{m}(\cdot), w(\cdot))$, we

also have $(\tilde{m}^\delta(\cdot), w(\cdot), \tilde{x}^\delta(\cdot), \tilde{\tau}^\delta) \Rightarrow (\tilde{m}(\cdot), w(\cdot), \tilde{x}(\cdot), \tilde{\tau})$, where (3.7) holds for the limit, and $\tilde{\tau}$ is the associated stopping time. Also $V(x, \tilde{m}^\delta) \to V(x, \tilde{m})$. We will use this fact to help approximate the $\tilde{m}(\cdot)$ policy so that it can be applied to $\xi^h(\cdot)$.

Next, given any $\rho > 0$, there is $\delta > 0$ such that we can approximate $\tilde{m}(\cdot)$ by an ordinary admissible control $\bar{u}^\rho(\cdot)$ with the following properties: (a) $\bar{u}^\rho(\cdot)$ takes only finitely many values (denoted by U_ρ); (b) it is constant on the intervals $[i\delta, i\delta + \delta)$, $i = 0, 1, \cdots$; (c) letting $\bar{m}^\rho(\cdot)$ denote the relaxed control representation of $\bar{u}^\rho(\cdot)$, we have

$$(\bar{m}^\rho(\cdot), \bar{x}^\rho(\cdot), w(\cdot), \bar{\tau}^\rho) \Rightarrow (\bar{m}(\cdot), \bar{x}(\cdot), w(\cdot), \bar{\tau}) \quad \text{as } \rho \to 0,$$

where $\bar{x}^\rho(\cdot)$ and $\bar{\tau}^\rho$ correspond to $(\bar{m}^\delta(\cdot), w(\cdot))$; (d) $V(x, \bar{m}^\rho) \leqq V(x) + \rho$. See [17] and [37] for the construction of such approximations. (Note that, under (A7.2) and the weak convergence, the hitting times of $\bar{x}^\rho(\cdot)$ on ∂G converge to those of $\bar{x}(\cdot)$ as $\rho \to 0$.)

We now prepare to choose a more appropriate control with properties (a)-(d), with possibly replacing ρ by 3ρ in (d). For each $\rho > 0$ and the δ of the last paragraph, consider an optimization problem for (3.7), (2.2), but where the controls are to be constant over the intervals $[p\delta, p\delta + \delta)$, $p = 0, 1, \cdots$, and take values in U_ρ. That is, a single value in U_ρ is used on each $[p\delta, p\delta + \delta)$. The optimization is *not* over the relaxed controls. This corresponds to controlling the discrete parameter Markov process obtained by sampling $x(\cdot)$ at times $p\delta$, $p = 0, 1, 2, \cdots$. (The optimal control for this "sampled" problem is an ordinary feedback control.) Let $\hat{u}^\rho(\cdot)$ denote the optimal control and $\hat{m}^\rho(\cdot)$ its relaxed control representation, and let $\hat{x}^\rho(\cdot)$ be the associated solution process. Since $\hat{m}^\rho(\cdot)$ is optimal in the chosen class of controls, we must have

$$V(x, \hat{m}^\rho) \leqq V(x) + \rho. \tag{7.6}$$

We next approximate $\hat{u}^\rho(\cdot)$ by a suitable function of $w(\cdot)$.

We note that for each given integer p, there is a measurable function $F_p^\rho(\cdot)$ such that $\hat{u}^\rho(t) = F_p^\rho(w(s), s \leqq p\delta)$ on $[p\delta, p\delta + \delta)$. We next approximate $F_p^\rho(\cdot)$ by a function that depends only on the samples of $w(\cdot)$ at a finite number of time points. Let $\theta < \delta$ such that δ/θ is an integer. There are U_ρ-valued measurable functions $F_p^{\rho,\theta}(\cdot)$ (of $w(i\theta)$, $i\theta \leqq p\delta$) such that for each δ, p,

$$F_p^{\rho,\theta}(w(i\theta), i\theta \leqq p\delta) \equiv u_p^{\rho,\theta} \to \hat{u}^\rho(p\delta)$$

w.p.1 as $\theta \to 0$. Let $m^{\rho,\theta}(\cdot)$ denote the relaxed control representation of the ordinary control $u^{\rho,\theta}(\cdot)$, which takes values $u_p^{\rho,\theta}$ on $[p\delta, p\delta + \delta)$, and let $x^{\rho\theta}(\cdot)$ and $\tau^{\rho\theta}$ denote the associated solution and stopping time. Then, for small enough θ, we have

$$V(x, m^{\rho,\theta}) \leqq V(x, \hat{m}^\rho) + \rho. \tag{7.7}$$

In fact, we can select $F_p^{\rho,\theta}(\cdot)$ such that there are a finite number of disjoint hyper-rectangles (open, closed, or partly open) that cover the range of its arguments such that $F_p^{\rho,\theta}(\cdot)$ is constant on each hyper-rectangle. Let us assume this form, and note that the probability is zero that the values of the random variable $\{w(i\theta), i\theta \leqq p\delta\}$ fall on the boundary of any hyper-rectangle.

We now adapt $F_p^{\rho,\theta}(\cdot)$ such that it can be applied to $\{\xi_n^h\}$. Recall the definition of $W^h(\cdot)$ given above Theorem 4.7. For n such that $p\delta \leqq t_n^h < p\delta + \delta$, use the control $F_p^{\rho,\theta}(W^h(i\theta), i\theta \leqq p\delta) \equiv \tilde{u}_n^h$. Let $\tilde{m}^h(\cdot)$ denote the relaxed control representation of the continuous parameter interpolation of $\{\tilde{u}_n^h\}$ (interpolation intervals $\Delta t_n^h = \Delta t^h(\xi_n^h, \tilde{u}_n^h)$). Then

$$\begin{aligned}(\xi^h(\cdot), \tilde{m}^h(\cdot), W^h(\cdot), \tau_h, F_p^{\rho,\theta}(W^h(i\theta), i\theta \leqq p\delta), p = 0, 1, \cdots) \\ \Rightarrow (x^{\rho\theta}(\cdot), m^{\rho,\theta}(\cdot), w(\cdot), \tau^{\rho\theta}, F_p^{\rho,\theta}(w(i\theta), i\theta \leqq p\delta), \quad p = 0, 1, \cdots).\end{aligned}$$

By the optimality of $V^h(x)$ and the above weak convergence,

$$V^h(x) \leqq V^h(x, \tilde{m}^h) \to V(x, m^{\rho,\theta}).$$

The above inequalities and convergence yield that $\overline{\lim}_h V^h(x) \leqq V(x) + 2\rho$ for the chosen subsequence. Since ρ is arbitrary and any subsequence of $\{\xi^h(\cdot), W^h(\cdot), m^h(\cdot), \tau_h\}$ has a subsequence that converges weakly, (7.5) holds. □

8. Controlled variance and drift. If $\sigma(x)$ is replaced by the controlled $\sigma(x, c)$ in (2.1) or (3.7), then the operator of the controlled diffusion is given by

$$\begin{aligned} \mathscr{L}^m f(x) &= \int f'_x(x) b(x, c) m_t(dc) + \frac{1}{2} \sum_{i,j} \int f_{x_i x_j}(x) a_{ij}(x, c) m_t(dc) \\ &= \int \mathscr{L}^c f(x) m_t(dc). \end{aligned} \tag{8.1}$$

Suppose for the moment that $U = \{c_1, \cdots, c_N\}$, a finite set of points, Then there are $p_i(t) \geqq 0, \sum_i p_i(t) = 1$, such that $\int a(x, c) m_t(dc) = \sum_1^N a(x, c_i) p_i(t)$. We can then represent some process $x(\cdot)$ with operator (8.1) as follows. Let $w_1(\cdot), \cdots, w_N(\cdot)$ be mutually independent standard vector-valued Wiener processes and for $B \in U$ define

$$M(B \times [0, t]) = \sum_i \int_0^t (p_i(s))^{1/2} \, dw_i(s) I_{\{c_i \in B\}} \equiv M(B, t).$$

Then, the process $x(\cdot)$ defined by

$$dx = \sum_{i=1}^N b(x, c_i) p_i(t) \, dt + \int \sigma(x, c) M(dc \, dt) \tag{8.2}$$

has the differential operator (8.1). The process $M(\cdot)$ can be viewed as a *measure-valued martingale*, and such processes provide a very useful basis for the representation and study of processes with operators (8.1). We now discuss (8.2) in a somewhat more general setting. For convenience in the notation we write $M(B \times [0, t]) = M(B, t)$.

Martingale measures. More detail as well as proofs of the assertions below concerning martingale measures and associated stochastic differential equations can be found in [14] and [51]. Let $\mathscr{F}_t$ be a filtration on some probability space, and let $\mathscr{U}$ denote the Borel sets of U. Let $M(\cdot)$ be a real-valued random function on $\Omega \times [0, \infty) \times \mathscr{U}$. We say that $M(\cdot)$ is a *measure-valued $\mathscr{F}_t$-martingale* or an *$\mathscr{F}_t$-martingale measure* with values $M(B, t)$ if $M(B, \cdot)$ is an $\mathscr{F}_t$-martingale for each $B \in \mathscr{U}$, and for each t, the following hold: $\sup_{B \in \mathscr{U}} EM^2(B, t) < \infty$, $M(A \cup B, t) = M(A, t) + M(B, t)$ w.p.1 for all disjoint $A, B \in \mathscr{U}$ and $EM^2(B_n, t) \to 0$ if $B_n \to \varnothing$, the empty set. Under (A8.1) below, $\sup_{B \in \mathscr{U}} EM^2(B, t) \leqq EM^2(U, t)$. If the $\mathscr{F}_t$ is unimportant or obvious, we omit it. $M(\cdot)$ is said to be *continuous* (respectively, square integrable) if each $M(B, \cdot)$ is. We say that $M(\cdot)$ is *orthogonal* if $M(A, \cdot)M(B, \cdot)$ is an $\mathscr{F}_t$-martingale whenever $A \cap B = \varnothing$. If $M(\cdot)$ and $N(\cdot)$ are $\mathscr{F}_t$-martingale measures and $M(A, \cdot)N(B, \cdot)$ is an $\mathscr{F}_t$-martingale for all Borel A, B, then $M(\cdot)$ and $N(\cdot)$ are said to be *strongly orthogonal.*

Let $M(\cdot) = (M_1(\cdot), \cdots, M_d(\cdot))'$, a vector-valued martingale measure. We henceforth suppose that

(A8.1) $M(\cdot) = (M_1(\cdot), \cdots, M_d(\cdot))'$ is square integrable and continuous, each component is orthogonal, and the pairs are strongly orthogonal.

By (A8.1), there are measure-valued (the values are measures on the Borel subsets of $U\times[0,\infty)$) random processes $m_i(\cdot)$ such that the quadratic variation processes satisfy, for each t and $B\in\mathcal{U}$,

$$\langle M_i(B,\cdot), M_j(A,\cdot)\rangle(t)=\delta_{ij}m_i(A\cap B,t)$$

where we write $m_i(A,t)$ for $m_i(A\times[0,t])$, the measure of $A\times[0,t]$.

We henceforth assume that (this will be the case in our application, anyway)

(A8.2) The m_i do not depend on i (we refer to it as $m(\cdot)$) and $m(U,t)=t$, for all t.

Under (A8.1) and (A8.2), for each Borel B, we have a predictable "derivative" process $m_\cdot(B)$ such that $m_t(\cdot)$ is a random measure on $\mathcal{U}$ and $m(dc\,dt)=m_t(dc)\,dt$.

Stochastic integrals. The stochastic integral with respect to a real-valued martingale measure $M(\cdot)$ is defined essentially as for real-valued martingales. Let $\mathcal{P}$ denote the σ-algebra of predictable sets in $\Omega\times[0,\infty)$ [25], [14], and $\mathcal{P}\times\mathcal{U}$ the σ-algebra over the product sets. For $f(\cdot)$ being $\mathcal{P}\times\mathcal{U}$ measurable, define

$$\|f\|_{T,m}=\left[E\int\int_0^T f^2(c,t)m(dc\,dt)\right]^{1/2} \qquad \text{for all } T<\infty.$$

$$L_m^2=\{f:\|f\|_{T,m}<\infty\}$$

For a bounded $f(\cdot)\in L_m^2$ taking constant values $f_i(c)$ on the intervals $[0,t_1]$, $(t_i,t_{i+1}]$, $i>0$, where $t_{i+1}>t_i$, we define the stochastic integral by

$$\psi(t)=\int\int_0^t f(c,s)M(dc\,ds)\equiv\sum_i\int f_i(c)[M(dc,t_{i+1}\wedge t)-M(dc,t_i\wedge t)].$$

Note that $E|\psi(T)|^2=\|f\|_m^2$. Now extend the definition to all $f(\cdot)\in L_m^2$ in the usual way [14], [25], [51].

The martingale problem. Let $b(\cdot,\cdot)$, $\sigma(\cdot,\cdot)$ be bounded and continuous and define $a=\sigma\sigma'$. Let there be a continuous process $x(\cdot)$ and a measure $m(\cdot)$ satisfying (A8.2) and such that for each bounded and smooth function $f(\cdot)$,

$$f(x(t))-f(x(0))-\int\int_0^t \mathcal{L}^c f(x)m_s(dc)\,ds\equiv Q_f(t)$$

is an $\mathcal{F}$-martingale, where $\mathcal{F}_t$ measures at least $\{x(s),m_s(\cdot),s\leqq t\}$. Then we say that $(x(\cdot),m(\cdot))$ solves the martingale problem for operator $\mathcal{L}^c$. Also [14] (possibly having to augment the probability space via the addition of Wiener processes or a martingale measure, which are independent of $m(\cdot)$, $x(\cdot)$), there is a martingale measure $M(\cdot)$ with quadratic variation $m(\cdot)I$ and satisfying (A8.1), (A8.2), and such that

$$dx=\int b(x,c)m_t(dc)\,dt+\int\sigma(x,c)M(dc\,dt). \tag{8.3}$$

Under a Lipschitz condition, we can say more. Suppose that

(A8.3) $b(\cdot)$, $\sigma(\cdot)$ are continuous, $b(\cdot,c)$, $\sigma(\cdot,c)$ are Lipschitz continuous uniformly in c and are bounded.

Assume (A8.3). Given $(M(\cdot),m(\cdot))$ satisfying (A8.1), (A8.2), with $\langle M(\cdot)\rangle=m(\cdot)I$, there is a unique strong-sense solution to (8.3) (which can be constructed by the classical "Picard iteration" technique). There is also a unique weak-sense solution in the sense that if $(M'(\cdot),m'(\cdot))$ and $(M(\cdot),m(\cdot))$ satisfy (A8.1), (A8.2), and have the same probability law, then the solution triples $(x'(\cdot),M'(\cdot),m'(\cdot))$,

$(x(\cdot), M(\cdot), m(\cdot))$ also have the same probability law. If $a(x, c)$ is uniformly (in $x \in G$, $c \in U$) positive definite, then (A7.2) holds under the "cone condition" described below it.

Admissible relaxed control. The system (8.3) represents our control system. It will be the representation of the limits of $\{\xi^h(\cdot)\}$ when the variance is also controlled. We say that $(M(\cdot), m(\cdot))$ is an *admissible relaxed control* for (8.3) if (A8.1) and (A8.2) hold and $\langle M(\cdot)\rangle = m(\cdot)I$. We continue to write the cost (3.9) as $V(x, m)$, even though its value depends on the joint distribution of $m(\cdot)$ and $M(\cdot)$.

Approximation of $(x(\cdot), M(\cdot), m(\cdot))$. Under (A8.1)-(A8.3), any such triple can be approximated by a triple satisfying

$$x^\delta(t) = x + \int_0^t \int b(x^\delta(s), c) m_s^\delta(dc)\, ds + \int_0^t \int \sigma(x^\delta(s), c) M^\delta(dc\, ds), \tag{8.4}$$

where $M^\delta(\cdot)$ is representable in terms of a finite number of Wiener processes. To get the approximation, let $\delta > 0$ and let $\{C_i^\delta, i \leqq k_\delta\}$ be a finite partition of U such that the diameters of $C_i^\delta \to 0$ as $\delta \to 0$. Let $c_i^\delta \in C_i^\delta$. Then $\{M(C_i^\delta, \cdot), i \leqq k_\delta\}$ are orthogonal continuous martingales with $\langle M(C_i^\delta, \cdot)\rangle = m(C_i^\delta, \cdot)$. There are mutually independent $\mathcal{F}_t$-standard Wiener processes $w_i^\delta(\cdot)$, $i \leqq k_\delta$, such that

$$M(C_i^\delta, t) = \int_0^t [m_s(C_i^\delta)]^{1/2}\, dw_i^\delta(s).$$

Let $M^\delta(\cdot)$ and $m^\delta(\cdot)$ be the restrictions of the measures $M(\cdot)$ and $m(\cdot)$, respectively, to the sets $\{C_i^\delta, i \leqq k_\delta\}$. Define $b_\delta(x, c) = b(x, c_i^\delta)$ and $\sigma_\delta(x, c) = \sigma(x, c_i^\delta)$, for $c \in C_i^\delta$, and define $x^\delta(\cdot)$ by

$$\begin{aligned} x^\delta(t) &= x + \int_0^t \int b_\delta(x^\delta(s), c) m_s^\delta(dc)\, ds + \int_0^t \int \sigma_\delta(x^\delta(s), c) M^\delta(dc\, ds) \\ &= x + \int_0^t \sum_i b(x^\delta(s), c_i^\delta) m_s(C_i^\delta)\, ds + \int_0^t \sum_i \sigma(x^\delta(s), c_i^\delta)[m_s(C_i^\delta)]^{1/2}\, dw_i^\delta(s). \end{aligned} \tag{8.5}$$

The $(x^\delta(\cdot), M^\delta(\cdot), m^\delta(\cdot))$ in (8.5) is the desired approximation, as shown by the following theorem.

THEOREM 8.1. *Assume* (A8.1)-(A8.3), *and define* $x^\delta(\cdot)$ *by* (8.5). *Then* $(x^\delta(\cdot), M^\delta(\cdot), m^\delta(\cdot)) \Rightarrow (x(\cdot), M(\cdot), m(\cdot))$. *Also* $V(x, m^\delta) \to V(x, m)$, *under* (A7.2) *or the random stopping rule of* § 7. *We can suppose that the* $m^\delta(\cdot)$ *is constant on intervals* $[p\delta, p\delta + \delta)$.

Proof. First, use the construction leading to (8.5), and ignore the last sentence of the theorem. By the construction, $(M^\delta(\cdot), m^\delta(\cdot)) \Rightarrow (M(\cdot), m(\cdot))$. We illustrate the proof of the convergence $x^\delta(\cdot) \to x(\cdot)$ only for a scalar case.

We have

$$\begin{aligned} x(t) - x^\delta(t) = &\int_0^t \int [b_\delta(x(s), c) - b_\delta(x^\delta(s), c)] m_s(dc)\, ds \\ &+ \int_0^t \int [\sigma_\delta(x(s), c) - \sigma_\delta(x^\delta(s), c)] M(dc\, ds) \\ &+ \int_0^t \int [b(x(s), c) - b_\delta(x(s), c)] m_s(dc)\, ds \end{aligned}$$

$$+\int_0^t \int [\sigma(x(s), c) - \sigma_\delta(x(s), c)]M(dc\, ds).$$

Now, use the facts that $|b(x, c) - b_\delta(x, c)| + |\sigma(x, c) - \sigma_\delta(x, c)|$ are bounded and go to zero as $\delta \to 0$ for each (x, c), together with the Lipschitz condition and the fact that $m_s(U) = 1$ and the properties of the martingales, to get that there is a $K < \infty$ such that

$$\sup_{\substack{t \leqq T \\ \delta}} (E|x(t)|^2 + E|x^\delta(t)|^2) < \infty,$$

$$E|x(s) - x^\delta(s)|^2 \leqq KE \int_0^t |x(s) - x^\delta(s)|^2\, ds + e^\delta(t),$$

$$E \sup_{s \leqq T} |x(s) - x^\delta(s)|^2 \leqq K \int_0^t E|x(s) - x^\delta(s)|^2\, ds + \tilde{e}^\delta(t),$$

where $e^\delta(\cdot)$ and $\tilde{e}^\delta(\cdot) \to 0$ as $\delta \to 0$ on each bounded interval. Thus $E \max_{s \leqq T} |x(s) - x^\delta(s)|^2 \to 0$ as $\delta \to 0$.

To get the last sentence of the theorem, we apply Theorem 3.1 to the system (8.5). □

The limit of the costs $V^h(x)$. The transition probabilities for the approximating chain can be calculated by the methods of § 5, simply by taking the control dependence of $a(x, c)$ into account. Define $V^h(x)$ by (7.1) again. We now proceed to characterize the limits.

THEOREM 8.2. *Assume* (2.4), (A2.2), *and* (A8.3). *Then* $\{\xi^h(\cdot), m^h(\cdot)\}$ *is*[2] *tight. If* $(x(\cdot), m(\cdot))$ *is the limit of a weakly convergent subsequence, then* $(x(\cdot), m(\cdot))$ *solves the martingale problem for operator* $\mathscr{L}^c$. *There is a martingale measure* $M(\cdot)$ *such that* $(M(\cdot), m(\cdot))$ *satisfy* (A8.1), (A8.2) *and* (8.3) *holds. Under the additional condition* (A7.2) *or the random stopping rule of* § 7, $V^h(x) \to V(x, m) \geqq V(x)$.

Remarks on the proof. The tightness proof is the same as in Theorem 4.5 or 4.6. The existence of $M(\cdot)$ was commented on above, and the use of (A7.2) is the same here as in § 7, to get convergence of the stopping times. We comment further only on the identification of the limit operator. The idea is to show (4.6) but with the definition of $\mathscr{L}^c$ incuding the c-dependence of $\sigma(x, c)$. But this follows from the proof of Theorem 4.6, once we note that all the expressions remain the same with control dependence added—provided that we write (we do the scalar case here, as in Theorem 4.6, for notational simplicity)

$$E_n^h f(\xi_{n+1}^h) - f(\xi_n^h) = \int f_x(\xi_n^h) b(\xi_n^h, c) m_n^h(dc) \Delta t_n^h$$

$$+ \frac{1}{2} f_{xx}(\xi_n^h) \int \sigma^2(\xi_n^h, c) m_n^h(dc) \Delta t_n^h + O(h^\alpha \Delta t_n^h)$$

$$= \int \mathscr{L}^c f(\xi_n^h) m_n^h(dc) \Delta t_n^h + O(h^\alpha \Delta t_n^h).$$

The convergence of the costs. By Theorem 8.1, we can approximate any admissible pair $(M(\cdot), m(\cdot))$ by a pair $(\tilde{M}(\cdot), \tilde{m}(\cdot))$, where $\tilde{m}(\cdot)$ is piecewise constant and takes finitely many values and where $\tilde{M}(\cdot)$ is represented in terms of a finite number of

[2] We use the terminology of § 7, where $m^h(\cdot)$ is the relaxed control representation of the continuous parameter interpolation (interpolation intervals $\{\Delta t_n^h = \Delta t^h(\xi_n^h, u_n^h)\}$) of $\{u_n^h\}$, the optimal control for $\{\xi_n^h\}$, under $\xi_0^h = x$.

Wiener processes. Thus, to prove that $V^h(x) \to V(x)$, we need only show that for any such pair

$$\overline{\lim_h} V^h(x) \leqq V(x, \tilde{m}). \tag{8.6}$$

THEOREM 8.3. *Under all the conditions of Theorem* 8.2, $V^h(x) \to V(x)$.

Proof. By the discussion above the theorem, we can suppose that $(\tilde{M}(\cdot), \tilde{m}(\cdot))$ is used and takes the following form. The $\tilde{m}_t(\cdot)$ are $\mathcal{F}_t$-adapted and piecewise constant, and are concentrated on the points $c_1, \cdots, c_q$, for all t. They are concentrated on one value of c_i on each interval $[p\delta, p\delta+\delta)$. The $w_i(\cdot)$, $i \leqq q$, are mutually independent $\mathcal{F}_t$-Wiener processes. Thus $\tilde{m}(\cdot)$ corresponds to an ordinary control $\tilde{u}(\cdot)$, where $\tilde{u}(\cdot)$ is constant on the intervals $[p\delta, p\delta+\delta)$ and takes values in the set $(c_i, \cdots, c_q)$. The model is

$$\begin{aligned} dx &= b(x, \tilde{u})\, dt + \sum_i \sigma(x, c_i) I_{\{\tilde{u}=c_i\}} dw_i \\ &= \sum_i b(x, c_i)\tilde{m}_t(c_i)\, dt + \sum_i \sigma(x, c_i)\tilde{m}_t^{1/2}(c_i)\, dw_i, \\ \tilde{m}_t(c_i) &= I_{\{\tilde{u}(t)=c_i\}}. \end{aligned} \tag{8.7}$$

Analogously to what was done in Theorem 7.1, we need only consider $\tilde{u}(\cdot)$, which are given by the functions $F_p^{\rho,\theta}(\cdot)$ introduced below (7.7) in that theorem. Here the arguments of these functions are the samples of all the Wiener processes $w_1(\cdot), \cdots, w_q(\cdot)$. We now construct the appropriate analogues of the $W^h(\cdot)$ used in Theorems 4.7 and 7.1.

As in Theorem 7.1, for convenience we suppose that σ is an $r \times r$-matrix. Let $\tilde{u}_n^h$ denote the ordinary admissible control to be used for the approximating chain $\{\xi_n^h\}$ (to be specified below). Let $\psi_i(\cdot)$, $i \leqq q$, be R^r-valued standard and mutually independent vector-valued Wiener processes, which are also independent of $\{\xi_n^h, \tilde{u}_n^h\}$, and define $\delta\psi_{i,n}^h = \psi_i(t_{n+1}^h) - \psi_i(t_n^h)$. Let $\tilde{m}_n^h$ and $\tilde{m}^h(\cdot)$ denote the relaxed control representation of $\tilde{u}_n^h$ and, respectively, the continuous parameter interpolation with derivative defined by $\tilde{m}_t^h = \tilde{m}_n^h$ on $[t_n^h, t_{n+1}^h)$. Recall the definition of δW_n^h above Theorem 4.7. Define the random variables

$$\delta W_{i,n}^h = \delta W_n^h I_{\{\tilde{u}_n^h = c_i\}} + \delta\psi_{i,n}^h I_{\{\tilde{u}_n^h \neq c_i\}}, \tag{8.8}$$

and define $W_i^h(t) = \sum_{t_{n+1}^h \leqq t} \delta W_{i,n}^h$.

Then we can write (where the process $\sum_{t_n^h \leqq t} \varepsilon_n^h = \varepsilon^h(t)$ goes to the zero process weakly as $h \to 0$):

$$\begin{aligned} \xi_{n+1}^h &= \xi_n^h + b(\xi_n^h, \tilde{u}_n^h)\Delta t_n^h + \sum_i \sigma(\xi_n^h, \tilde{u}_n^h) I_{\{\tilde{u}_n^h = c_i\}} \delta W_{i,n}^h + \varepsilon_n^h \\ &= \xi_n^h + \sum_i b(\xi_n^h, c_i)\tilde{m}_n^h(c_i)\Delta t_n^h + \sum_i \sigma(\xi_n^h, c_i)|\tilde{m}_n^h(c_i)|^{1/2} \delta W_{i,n}^h + \varepsilon_n^h, \\ \tilde{m}_n^h(c_i) &= I_{\{\tilde{u}_n^h = c_i\}}. \end{aligned} \tag{8.9}$$

We continue to follow the procedure of Theorem 7.1. For n such that $t_n^h < \delta$, use any control. For $p = 1, 2, \cdots$ and n such that $t_n^h \in [p\delta, p\delta+\delta)$, use the control defined by $\tilde{u}_n^h = F_p^{\rho\theta}(W_i^h(j\theta), j \leqq p\delta/\theta, i \leqq q)$. As in Theorem 7.1, this yields

$$(\xi^h(\cdot), W_i^h(\cdot), i \leqq q, \tilde{m}^h(\cdot), \tilde{\tau}_h) \Rightarrow (\tilde{x}(\cdot), w_i(\cdot), i \leqq q, \tilde{m}(\cdot), \tilde{\tau}),$$

where the limit satisfies (8.7) and $\tilde{\tau}$ and $\tilde{\tau}^h$ are the escape times from G. Hence, since $V^h(x)$ is the infimum cost, we have

$$V^h(x) \leqq V^h(x, \tilde{m}^h) \to V(x, \tilde{m}). \tag{8.10}$$

□

9. Control of queueing and production systems in heavy traffic. There has been a considerable amount of work done on the heavy traffic modeling of queueing and production systems [22], [31], [39], [44], but rather little on the use of such modeling for control purposes [31]. Since this class of problems is of increasing importance and little is available concerning the numerical problem, this section is particularly timely. It also gives us an illustration of how to treat an important class of reflection problems where the reflection is defined in terms of a discrete physical problem, and is discontinuous. In this section, we discuss one simple case in order to illustrate the applicability of the "Markov chain" approximation for computational purposes for such problems. We concentrate on the two-dimensional process illustrated in Fig. 9.1, although the technique and results hold true, in general, and the procedure that we follow should indicate further possibilities. A somewhat different control problem is developed and a Markov chain approximation discussed in [31]. Further work on other formulations involving controlled routing and singular controls will appear in a forthcoming paper.

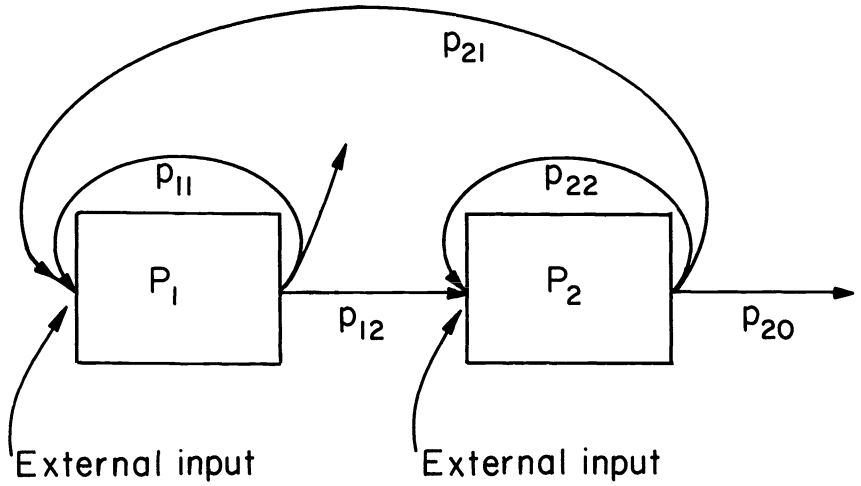


FIG. 9.1. *The system model.*

By heavy traffic, we mean essentially that the processors have little idle time. ε is a measure of the idle time (made more precise in (A9.1), (A9.2)). As $\varepsilon \to 0$, the idle time goes to zero and the traffic intensity goes to one. The scaling used below allows the physical system to be approximated by a reflected diffusion, and allows a very complicated system to have a relatively simple approximation.

In this section the development is somewhat different from that in other sections. Here the physical model is not given a priori as some sort of controlled diffusion as it was in the previous sections, but it is given in the "physical" form as a pair of interacting processors with various (discrete parameter and asynchronized) inputs and outputs. We need *first* to find the proper (heavy traffic) limit model, and *then* apply a Markov chain numerical approximation method to this limit model. Generally, the physical model is much too complicated for us to have any reasonable hope of solving an optimization problem for it directly. But for the numerical approximation to be valid, we must show that (a) the limit of the optimal costs for the Markov chain approximation is the optimal cost for the controlled heavy traffic limit and (b), the optimal cost for the latter problem is the limit of the optimal costs for the physical models as $\varepsilon \to 0$. Thus, we are obliged first to find the correct limit controlled model—for it is not a priori obvious. (The precise control problem will be specified below.) Part

of the development of the limit of the "physical" model parallels work that would be necessary to obtain convergence of the Markov chain approximation to the limit model.

For our problem, there are two interconnected processors P_1 and P_2. The outputs of the processors are routed according to the arrows in Fig. 9.1, with p_{ij} denoting the probability that an output of processor i goes to processor j; p_{i0} denotes the probability that the serviced object leaves the system. Both processors have external inputs. We will "marginally" control the distribution of the interarrival and service intervals. This "marginal" control will have a substantial effect in the heavy traffic case. Since we are working with a "physical" system, which is described in terms of inputs and outputs (or service and interarrival intervals and routing), we must introduce some notation, which allows us to describe the physical system in a way that allows a convenient derivation of the limit. Owing to the control and the state dependencies, we are forced to deviate from the classical approach, which might be best represented by [44].

DEFINITIONS. For $\varepsilon > 0$ and $i = 1, 2$, let $\{\alpha_n^{i,\varepsilon}, n < \infty\}$, $\{\Delta_n^{i,\varepsilon}, n < \infty\}$, respectively, denote the sequence of interarrival intervals for the external inputs to P_i and the service intervals of P_i, respectively. For notational convenience, we let the physical model evolve in discrete time. The buffer size of P_i is $B_i/\sqrt{\varepsilon}$, assumed to be an integer. Define

$$S_{a,n}^{i,\varepsilon} = \sum_1^n \alpha_j^{i,\varepsilon}, \qquad S_{d,n}^{i,\varepsilon} = \sum_1^n \Delta_j^{i,\varepsilon}.$$

Let $\mathcal{F}_{a,n}^{i,\varepsilon}$ be the σ-algebra determined by all service and arrival intervals and routings that are completed by (discrete) time $S_{a,n}^{i,\varepsilon}$, the time of the nth external arrival to P_i, as well as all the intervals other than $\alpha_{n+1}^{i,\varepsilon}$ that start at or before $S_{a,n}^{i,\varepsilon}$. Let $E_{a,n}^{i,\varepsilon}$ be the expectation conditioned on $\mathcal{F}_{a,n}^{i,\varepsilon}$. Define $\mathcal{F}_{d,n}^{i,\varepsilon}$ and $E_{d,n}^{i,\varepsilon}$ analogously, and analogously define the conditional variances $\text{var}_{a,n}^{i,\varepsilon}$ and $\text{var}_{d,n}^{i,\varepsilon}$. We use $E_{a,n,c}^{i,\varepsilon}$ to denote the conditional expectation given (in addition to the data used above) that control value c was applied to the $(n+1)$st interarrival interval, and analogously define $E_{d,n,c}^{i,\varepsilon}$. Let $\xi_n^{i,\varepsilon}$ (respectively, $\psi_n^{i,\varepsilon}$) denote the indicator function of an external arrival to P_i (respectively, a service completion at P_i) at time n, and let $I_n^{ij,\varepsilon}$ denote the indicator function of a routing from P_i to P_j if a service is completed at time n. $I_n^{i0,\varepsilon} = 1$ denotes that the routing was to the "exterior."

Let $Q_n^{i,\varepsilon}$ denote the number of items in or waiting for service at P_i at time n. Following the usual scaling used in heavy traffic limit theorems [31], [44], [39], set $X_n^{i,\varepsilon} = \sqrt{\varepsilon}\, Q_n^{i,\varepsilon}$, $X^{i,\varepsilon}(t) = \sqrt{\varepsilon} Q_{t/\varepsilon}^{i,\varepsilon}$. The ratio t/ε is always used to denote the integer part. With this notation, we have $X^{i,\varepsilon}(\varepsilon S_{a,n}^{i,\varepsilon}) = X_{S_{a,n}^{i,\varepsilon}}^{i,\varepsilon}$. Write $X = (X^1, X^2)$. Thus $0 \leqq X_n^{i,\varepsilon} \leqq B_i$.

We next state some of the heavy traffic assumptions. "Heavy traffic" means that the processors have very little idle time; i.e., the mean rates of arrival and service are very close for each server. We quantify this difference by a factor of order $\sqrt{\varepsilon}$ in the mean rates. It is also supposed that the mean arrival or mean service intervals can be controlled. But the control here will have only a marginal effect on the rates, although a major effect on the limit controlled diffusion. In situations where heavy traffic modeling is appropriate, these marginal effects on the rates have substantial effects on system variables such as mean queue lengths and the probability of a full buffer (thus denying entrance to a new arrival).

We suppose that the control takes values in a set $U = U_{a1} \times U_{d1} \times U_{a2} \times U_{d2}$, where the U_α are compact; the set of control values to be applied to control the interarrival time to P_i is U_{ai}, etc. Thus each "activity" can be controlled separately. We will use the following assumptions. They are simply the "control" forms of analogous assumptions in, e.g., [44].

(A9.0) The sequence of routings are mutually independent and independent of $\{\alpha_n^{i,\varepsilon}, \Delta_n^{i,\varepsilon}\}$. The spectral radius of $P=\{p_{ij}\}$ is less than unity. Completed services at P_i, which are routed back to P_i, have priority.

(A9.1) There are constants $\{g_\alpha\}$ and bounded and continuous real-valued functions $a^i(\cdot)$, $d^i(\cdot)$, $\sigma_a^i(\cdot)$, $\sigma_d^i(\cdot)$, $i=1,2$, and $\delta_\varepsilon, \delta'_\varepsilon \to 0$, such that for all x, if $X^\varepsilon_{S^{i,\varepsilon}_{a,n}} = x$, then

$$E^{i,\varepsilon}_{a,n,c}\alpha^{i,\varepsilon}_{n+1} \equiv \bar{\alpha}^{i,\varepsilon}_{n+1} = [g_{ai} + \sqrt{\varepsilon}\, a^i(x, c) + o(\sqrt{\varepsilon})]^{-1},$$
$$\operatorname{var}^{i,\varepsilon}_{a,n,c}\alpha^{i,\varepsilon}_{n+1} = [\sigma_a^i(x)]^2 + \delta_\varepsilon.$$

Also, for all x, if $X^\varepsilon_{S^{i,\varepsilon}_{d,n}} = x$, then

$$E^{i,\varepsilon}_{d,n,c}\Delta^{i,\varepsilon}_{n+1} \equiv \bar{\Delta}^{i,\varepsilon}_{n+1} = [g_{di} + \sqrt{\varepsilon}\, d^i(x, c) + o(\sqrt{\varepsilon})]^{-1},$$
$$\operatorname{var}^{i,\varepsilon}_{d,n,c}\Delta^{i,\varepsilon}_{n+1} = [\sigma_d^i(x)]^2 + \delta'_\varepsilon.$$

Define $g_a = (g_{a1}, g_{a2})'$, $g_d = (g_{d1}, g_{d2})'$. Assumption (A9.2) says that the mean arrival and departure rates for each P_i are equal (modulo $O(\sqrt{\varepsilon})$), and is the commonly used heavy traffic assumption [22], [31], [44], [39].

(A9.2) $g_a = (I - P')g_d$.

(A9.3) $\{|\alpha_n^{i,\varepsilon}|^2, |\Delta_n^{i,\varepsilon}|^2,\ i,\ n,$ small ε, all controls$\}$ is uniformly integrable.

Admissible controls. The *admissible controls* for the nth service or interarrival intervals at each P_i are (U_{di} or U_{ai}-valued) functions only of the data available up to the start of those intervals.

We will use the following terminology for the controls. As in the previous sections, the actual control used on the physical process (or on the Markov chain approximation) will be an ordinary and not a relaxed control. But, as before, the relaxed control terminology is useful for getting the appropriate limits. Let $U^{i,\varepsilon}_{a,k}$ denote the actual ordinary control used to control the kth interarrival interval for the external inputs to P_i. For discrete time n, define $u^{i,\varepsilon}_{a,n} = U^{i,\varepsilon}_{a,k}$ for $n \in [S^{i,\varepsilon}_{a,k}, S^{i,\varepsilon}_{a,k+1})$. Let $m^{i,\varepsilon}_{a,n}(\cdot)$ denote the relaxed control representation of $u^{i,\varepsilon}_{a,n}$ (i.e., $m^{i,\varepsilon}_{a,n}(C) = I_{\{u^{i,\varepsilon}_{a,n} \in C\}}$). Let $m^{i,\varepsilon}_a(\cdot)$ denote the continuous parameter piecewise constant interpolation defined by the derivative: $m^{i,\varepsilon}_{a,t}(\cdot) = m^{i,\varepsilon}_{a,n}(\cdot)$ for $t \in [\varepsilon n, \varepsilon n + \varepsilon)$. Define $u^{i,\varepsilon}_{d,n}$ and $m^{i,\varepsilon}_d(\cdot)$ analogously, Let u^ε_n and $m^\varepsilon(\cdot)$ denote the vectors composed (each) of the four components $\{u^{i,\varepsilon}_{\alpha,n},\ i=1,2,\ \alpha = a, d\}$, $\{m^{i,\varepsilon}_\alpha(\cdot),\ i=1,2,\ \alpha = a, d\}$, respectively.

The cost criterion. In order to illustrate the possibilities, we choose a discounted cost of the following form. We wish to penalize both the cost of control, the holding or inventory cost, and the "opportunity cost" due to nonadmission of an arrival to the P_i due to a full buffer.

The (scaled by $\sqrt{\varepsilon}$) number of lost inputs due to a full buffer at P_i can be written as ($j \neq i$)

$$L^{i,\varepsilon}_n = \sqrt{\varepsilon} \sum_{k=0}^{n} \xi^{i,\varepsilon}_k I_{\{X^{i,\varepsilon}_k = B_i\}} + \sqrt{\varepsilon} \sum_{k=0}^{n} I^{ji,\varepsilon}_k \psi^{j,\varepsilon}_k I_{\{X^j_k \neq 0, X^j_k = B_i\}}. \tag{9.1}$$

Define $L^{i,\varepsilon}(t) = L^{i,\varepsilon}_{t/\varepsilon}$. For $\beta > 0$, the cost function is

$$V^\varepsilon(x, m) = E^m_x \int_0^\infty e^{-\beta t} \int k(X^\varepsilon(t), c) m_t(dc)\, dt + E^m_x \int_0^\infty e^{-\beta t}[k_1 dL^{1,\varepsilon}(t) + k_2 dL^{2,\varepsilon}(t)]. \tag{9.2}$$

We assume

(A9.4) $k(\cdot,\cdot)$ is bounded and continuous and $k_i > 0$.

The system equations. A classical problem in getting heavy traffic limit theorems concerns the situation when the buffers are empty. A common way of handling this is to assume that the processors keep processing, even if there is nothing to process. The consequent sequence of "phantom" outputs must be compensated for by suitable reflection terms. If an arrival occurs in the midst of a service interval in which a "phantom" object is being processed, then the actual service time for that arrival is taken to be the residual time of the current service interval [24], [44]. In classical cases [24], it can be proved that the modification of the $X^\varepsilon(\cdot)$ process is asymptotically insignificant. We use this "residual time" assumption here also, and (although we omit the proof), the asymptotic insignificance can also be shown. (See also the appendix of [31] for a related argument for another heavy traffic control problem.)

Define the "reflection" terms for $j \neq 0$,

$$Y_n^{ij,\varepsilon} = \sqrt{\varepsilon} \sum_{k=1}^{n} \psi_k^{i,\varepsilon} I_k^{ij,\varepsilon} I_{\{X_k^{i,\varepsilon}=0\}} \tag{9.3}$$

and set $Y^{ij,\varepsilon}(t) = Y_{t/\varepsilon}^{ij,\varepsilon}$. The $Y_n^{i,j,\varepsilon}$ is just the number (scaled by $\sqrt{\varepsilon}$) of "phantom" inputs from i to j that occurred by discrete time n.

In subsequent sums, we drop the ε-affix of the summands. We can write

$$\begin{aligned} X_n^{1,\varepsilon} &= X_0^{1,\varepsilon} + \sqrt{\varepsilon} \sum_{i}^{n} \xi_k^1 - \sqrt{\varepsilon} \sum_{1}^{n} \psi_k^1 (I_k^{12} + I_k^{10}) + \sqrt{\varepsilon} \sum_{1}^{n} \psi_k^2 I_k^{21} \\ &\quad + (Y_n^{12,\varepsilon} + Y_n^{10,\varepsilon}) - Y_n^{21,\varepsilon} - L_n^{1,\varepsilon}, \\ X_n^{2,\varepsilon} &= X_0^{2,\varepsilon} + \sqrt{\varepsilon} \sum_{1}^{n} \xi_k^2 - \sqrt{\varepsilon} \sum_{1}^{n} \psi_k^2 (I_k^{21} + I_k^{20}) + \sqrt{\varepsilon} \sum_{1}^{n} \psi_k^1 I_k^{12} \\ &\quad + (Y_n^{21,\varepsilon} + Y_n^{20,\varepsilon}) - Y_n^{12,\varepsilon} - L_n^{2,\varepsilon}. \end{aligned} \tag{9.4}$$

A useful form for system equations. We next go through a sequence of manipulations whose purpose is to put in a form that allows for a relatively straightforward weak convergence proof, although it does require some new definitions. The idea is to write the terms involving ψ or ξ as sums of scaled "martingales" and "drift" terms, which can be handled more easily than (9.4) can. We now center the second to fourth terms on the right of each equation in (9.4) so that they can be written as a sum of a martingale and a drift term. Define the random variables and processes:

$$\delta M_{a,k}^{i,\varepsilon} = \left[1 - \frac{\alpha_k^{i,\varepsilon}}{\bar{\alpha}_k^{i,\varepsilon}}\right], \qquad \delta M_{d,k}^{ij,\varepsilon} = \left[I_k^{ij,\varepsilon} - p_{ij} \frac{\Delta_k^{i,\varepsilon}}{\bar{\Delta}_k^{i,\varepsilon}}\right],$$

$$M_a^{i,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{1}^{t/\varepsilon} \delta M_{a,k}^{i}, \qquad M_d^{ij,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{1}^{t/\varepsilon} \delta M_{d,k}^{ij,\varepsilon}.$$

The subscript k in $I_k^{ij,\varepsilon}$ should have been $S_{d,k}^{i,\varepsilon}$, the time of the kth departure, but the replacement used yields the same results owing to the independence assumption on the routing variables. The $\delta M_{d,n}^{i,\varepsilon}$ and $\delta M_{a,n}^{ij,\varepsilon}$ are martingale differences with respect to the filtrations $\mathcal{F}_{d,n}^{i,\varepsilon}$ and $\mathcal{F}_{a,n}^{i,\varepsilon}$, respectively.

By a straightforward evaluation using the representation of the $\bar{\alpha}$ and $\bar{\Delta}$ in (A9.1), we get (letting $x = X^{\varepsilon}_{S^{i,\varepsilon}_{a,k}}$ or $X^{\varepsilon}_{S^{i,\varepsilon}_{d,k}}$, as appropriate)

$$\text{var}^{i,\varepsilon}_{a,k}\,\delta M^{i,\varepsilon}_{a,k+1} = g^2_{ai}(\sigma^i_a(x))^2 + O(\sqrt{\varepsilon}) + O(\delta_\varepsilon), \tag{9.5}$$

$$\begin{aligned} &E^{i,\varepsilon}_{d,k}[\delta M^{i0,\varepsilon}_{d,k+1} \cdot \delta M^{ij,\varepsilon}_{d,k+1}] = -p_{ij}p_{i0} + p_{ij}p_{i0}g^2_{di}(\sigma^i_d(x))^2 + O(\sqrt{\varepsilon}) + O(\delta_\varepsilon), \quad j \neq 0, i, \\ &\text{var}^{i,\varepsilon}_{d,k}\,[\delta M^{ij,\varepsilon}_{d,k+1}] = p_{ij}(1-p_{ij}) + g^2_{di}p^2_{ij}(\sigma^i_d(x))^2 + O(\sqrt{\varepsilon}) + O(\delta_\varepsilon). \end{aligned} \tag{9.6}$$

Define $\sigma^2_{ai}(x) = g^2_{ai}(\sigma^i_a(x))^2$, and define the random vectors and processes

$$\delta M^{1,\varepsilon}_{d,k} = (-(\delta M^{10,\varepsilon}_{d,k} + \delta M^{12,\varepsilon}_{d,k}),\ \delta M^{12,\varepsilon}_{d,k})',$$

$$\delta M^{2,\varepsilon}_{d,k} = (\delta M^{21,\varepsilon}_{d,k},\ -(\delta M^{20,\varepsilon}_{d,k} + \delta M^{21,\varepsilon}_{d,k}))',$$

$$M^{i,\varepsilon}_{d,n} = \sqrt{\varepsilon}\sum_1^n \delta M^{i,\varepsilon}_{d,k}, \qquad M^{i,\varepsilon}_d(t) = M^{i,\varepsilon}_{d,t/\varepsilon}.$$

From (9.6) we can calculate the matrix $\sum_{di}(x)$ defined by $\text{cov}^{i,\varepsilon}_{d,k}\,\delta M^{i,\varepsilon}_{d,k+1} = \sum_{di}(x) + O(\sqrt{\varepsilon}) + O(\delta\varepsilon)$.

Although the nondegeneracy in the next assumption (A9.5) is not needed for the validity of Theorem 9.1, it simplifies the notation a bit.

(A9.5) $\sum_{di}(x) > 0$, $\inf_x \sigma^2_{ai}(x) > 0$ and there are continuous $\sigma_{di}(x)$ such that $\sum_{di}(x) = \sigma_{di}(x)\sigma'_{di}(x)$ and $\sigma^{-1}_{di}(x)$ is uniformly bounded.

Let $S^{i,\varepsilon}_\alpha(t) = \varepsilon S^{i,\varepsilon}_{\alpha,t/\varepsilon}$ for $\alpha = a, d$, and define the process $\bar{S}^{i,\varepsilon}_a(t) = \max\{\varepsilon k : \varepsilon S^{i,\varepsilon}_{a,k} \leqq t\}$ and analogously define $\bar{S}^{i,\varepsilon}_d(t)$. The function $\bar{S}^{i,\varepsilon}_a(t)$ is an inverse of the function $S^{i,\varepsilon}_a(t)$, and is just ε times the number of external inputs to P_i, which arrive by discrete time t/ε (or interpolated time t).

We now rewrite (9.4) in a continuous parameter form from which it is easier to get the heavy traffic limit. By the definition of $\bar{S}^{i,\varepsilon}_a(\cdot)$ and $\delta M^{i,\varepsilon}_{a,k}$, we can write

$$\sqrt{\varepsilon}\sum_1^{t/\varepsilon}\xi^i_{a,k} = \sqrt{\varepsilon}\sum_{k=1}^{\varepsilon^{-1}\bar{S}^{1,\varepsilon}_a(t)}\delta M^i_{a,k} + \sqrt{\varepsilon}\sum_{k=1}^{\varepsilon^{-1}\bar{S}^{i,\varepsilon}_a(t)}\frac{\alpha^i_{a,k}}{\bar{\alpha}^i_{a,k}},$$

etc. Thus we can write

$$\begin{aligned} X^{1,\varepsilon}(t) &= X^{1,\varepsilon}(0) + M^{1,\varepsilon}_a(\bar{S}^{1,\varepsilon}_a(t)) - [M^{10,\varepsilon}_d(\bar{S}^{1,\varepsilon}_d(t)) + M^{12,\varepsilon}_d(\bar{S}^{1,\varepsilon}_d(t))] \\ &\quad + B^{1,\varepsilon}(t) + M^{21,\varepsilon}_d(\bar{S}^{2,\varepsilon}_d(t)) + (Y^{10,\varepsilon}(t) + Y^{12,\varepsilon}(t)) - Y^{21,\varepsilon}(t) - L^{1,\varepsilon}(t), \end{aligned} \tag{9.7}$$

where

$$B^{1,\varepsilon}(t) = \sqrt{\varepsilon}\sum_{k=1}^{\varepsilon^{-1}\bar{S}^{1,\varepsilon}_a(t)}\frac{\alpha^1_k}{\bar{\alpha}^1_k} - (p_{10}+p_{12})\sqrt{\varepsilon}\sum_{k=1}^{\varepsilon^{-1}\bar{S}^{1,\varepsilon}_d(t)}\frac{\Delta^1_k}{\bar{\Delta}^1_k} + p_{21}\sqrt{\varepsilon}\sum_{k=1}^{\varepsilon^{-1}\bar{S}^{2,\varepsilon}_d(t)}\frac{\Delta^2_k}{\bar{\Delta}^2_k}. \tag{9.8}$$

We get the expression for $X^{2,\varepsilon}(t)$ by just interchanging 1 and 2 in (9.7) and (9.8).

Equation (9.8) is readily simplified. Note that

$$\sqrt{\varepsilon}\sum_{k=1}^{\varepsilon^{-1}\bar{S}^{1,\varepsilon}_a(t)}\alpha^1_k = \frac{t}{\sqrt{\varepsilon}} + O(\sqrt{\varepsilon}),$$

with a similar expression for the other "bad $1/\sqrt{\varepsilon}$" terms in (9.8) involving the Δ^α_β. The heavy traffic assumption (A9.2) and the expansion of $\bar{\alpha}$ and $\bar{\Delta}$ in (A9.1) allows

the "bad $1/\sqrt{\varepsilon}$" terms in (9.8) to sum to zero mod $O(\sqrt{\varepsilon})$. Equation (9.8) then reduces to

$$B^{1,\varepsilon}(t) = \varepsilon \sum_{1}^{t/\varepsilon} b^1(X_k^\varepsilon, u_k^\varepsilon) + \delta_\varepsilon^1 = \int_0^t \int b^1(X^\varepsilon(s), c) m_s^\varepsilon(dc)\, ds + \delta_\varepsilon^1, \tag{9.9}$$

where $\delta_\varepsilon^1 \xrightarrow{\varepsilon} 0$ and

$$b^1(x, c) = a^1(x, c) - (p_{10} + p_{12}) d^1(x, c) + p_{21} d^2(x, c),$$

and similarly for $B^{2,\varepsilon}(t)$.

We can now write (9.7) in a better form. Define the scaled processes $\bar{M}_d^{i,\varepsilon}(t) = M_d^{i,\varepsilon}(\bar{S}_d^{i,\varepsilon}(t))$ and $\bar{M}_a^{i,\varepsilon}(t) = M_a^{i,\varepsilon}(\bar{S}_a^{i,\varepsilon}(t))$. Define $Y_n^{i,\varepsilon} = \sum_1^n \psi_k^i I_{\{X_k^{i,\varepsilon}=0\}}$, and let $Y^\varepsilon(\cdot) = (Y^{1,\varepsilon}(\cdot), Y^{2,\varepsilon}(\cdot))'$ denote the continuous parameter interpolation (interval ε). Define the "error terms" ($i \neq j$, $i, j = 1, 2$) $\tilde{Y}^{i,\varepsilon}(\cdot)$ and $\tilde{Y}^{ij,\varepsilon}(\cdot)$ by

$$\tilde{Y}^{i,\varepsilon}(t) = [Y^{i0,\varepsilon}(t) - p_{i0} Y^{i,\varepsilon}(t)] + [Y^{ij,\varepsilon}(t) - p_{ij} Y^{i,\varepsilon}(t)] \equiv \tilde{Y}^{i0,\varepsilon}(t) + \tilde{Y}^{ij,\varepsilon}(t).$$

Define $\tilde{Y}^\varepsilon(\cdot) = (\tilde{Y}^{1,\varepsilon}(\cdot), \tilde{Y}^{2,\varepsilon}(\cdot))'$. Then, finally, (9.7) can be written in the form that will be used in the next theorem. Let $x(0) = X^\varepsilon(0)$. Then

$$\begin{aligned} X^\varepsilon(t) &= x(0) + \begin{pmatrix} \bar{M}_a^{1,\varepsilon}(t) \\ \bar{M}_a^{2,\varepsilon}(t) \end{pmatrix} + \bar{M}_d^{1,\varepsilon}(t) + \bar{M}_d^{2,\varepsilon}(t) + B^\varepsilon(t) \\ &\quad + \tilde{Y}^\varepsilon(t) + (I - P') Y^\varepsilon(t) - L^\varepsilon(t) \\ &\equiv Z^\varepsilon(t) + (I - P') Y^\varepsilon(t) - L^\varepsilon(t). \end{aligned} \tag{9.10}$$

In order to do a "comparison control" argument like that in Theorem 7.1, a suitable replacement for the δW^h of Theorems 4.7 and 7.1 is needed. To prepare for that, we need the following definitions. We let x denote the value of the system state at the time of the kth arrival to or departure from P_i, as appropriate.

Define (using x as above (9.5))

$$\delta W_{a,k+1}^{i,\varepsilon} = \sqrt{\varepsilon}\, \delta M_{a,k+1}^{i,\varepsilon} / (\sqrt{g_{ai}}\, \sigma_{ai}(x)).$$

Then

$$\operatorname{var}_{a,k}^{i,\varepsilon} \delta W_{a,k+1}^{i,\varepsilon} = \varepsilon / g_{ai} + O(\varepsilon^{3/2}).$$

Analogously, define the "2-vectors" $\delta W_{d,k+1}^{i,\varepsilon} = \sqrt{\varepsilon}\, \sigma_{di}^{-1}(x) \delta M_{d,k+1}^{i,\varepsilon} / \sqrt{g_{di}}$. Define $W_{\alpha,n}^{i,\varepsilon} = \sum_0^{n-1} \delta W_{\alpha,k}^{i,\varepsilon}$, $\alpha = a, d$, and define $W_\alpha^{i,\varepsilon}(t) = W_{\alpha,t/\varepsilon}^{i,\varepsilon}$ and $\bar{W}_\alpha^{i,\varepsilon}(t) = W_\alpha^{i,\varepsilon}(\bar{S}_\alpha^{i,\varepsilon}(t))$.

Heavy traffic limit theorems. We can now state the first heavy traffic limit theorem.

THEOREM 9.1. *Assume* (A9.0)-(A9.5). *Then*

$$R^\varepsilon(\cdot) = \{X^\varepsilon(\cdot), Y^\varepsilon(\cdot), \tilde{Y}^\varepsilon(\cdot), \bar{W}_\alpha^{i,\varepsilon}(\cdot), \bar{M}_\alpha^\varepsilon(\cdot), L^\varepsilon(\cdot), B^\varepsilon(\cdot), m^\varepsilon(\cdot), i = 1, 2, \alpha = a, d\}$$

is tight (*all components in the Skorokhod topology on* $D^k[0, \infty)$ *for the appropriate* k, *except for* $m^\varepsilon(\cdot)$). *If*

$$R(\cdot) = \{x(\cdot), Y(\cdot), \tilde{Y}(\cdot), \bar{w}_\alpha^i(\cdot), \bar{M}_\alpha(\cdot), L(\cdot), B(\cdot), m(\cdot), i = 1, 2, \alpha = a, d\}$$

denotes the limit of a weakly convergent subsequence, then $\tilde{Y}(\cdot) =$ *zero process, all limits are continuous and*

$$x(t) = x(0) + \bar{M}_a(t) + \bar{M}_d^1(t) + \bar{M}_d^2(t) + (I - P') Y(t) + B(t) - L(t), \tag{9.11}$$

where the $Y^i(\cdot)$ (*respectively,* $L^i(\cdot)$) *are nonnegative, nondecreasing and can increase only when* $X^i(t) = 0$ (*respectively,* $X^i(t) = B_i$). *Let* $\mathcal{R}_t$ *denote the* σ*-algebra generated by*

$\{R(u), u \leqq t\}$. *Then the* $\bar{w}_i^{\alpha}(\cdot)$ *are standard and mutually independent* $\mathcal{R}_t$*-Wiener processes. Furthermore (where* $m_t(\cdot)$ *denotes the time derivative of* $m(\cdot)$ *at* t*),*

$$B^i(t) = \int_0^t \int b^i(x(s), c) m_s(dc)\, ds. \tag{9.12}$$

Define $\bar{M}_a = (\bar{M}_a^1, \bar{M}_a^2)$. *The* $\bar{M}_{\alpha}^i(\cdot)$, $i = 1, 2$, $\alpha = a, d$, *are mutually orthogonal* $\mathcal{R}_t$*-martingales with*

$$\begin{aligned} &\text{quadratic variation } \bar{M}_a^i(t) = \langle \bar{M}_a^i \rangle(t) = g_{ai} \int_0^t \sigma_{ai}^2(x(s))\, ds, \\ &\text{quadratic variation } \bar{M}_d^i(t) = \langle \bar{M}_d^i \rangle(t) = g_{di} \int_0^t \sum_{di} (x(s))\, ds. \end{aligned} \tag{9.13}$$

They have the representation

$$\bar{M}_a^i(t) = g_{ai}^{1/2} \int_0^t \sigma_{ai}(x(s)) d\bar{w}_a^i(s), \qquad \bar{M}_d^i(t) = g_{di}^{1/2} \int_0^t \sigma_{di}(x(s)) d\bar{w}_d^i(s). \tag{9.14}$$

Also, $m(\cdot)$ *is nonanticipative with respect to* $\{\bar{w}_{\alpha}^i(\cdot), i = 1, 2, \alpha = a, d\}$.

Proof. The details are very similar to those for an impulsively controlled heavy traffic problem in [31, § 5], except that here we use a continuously acting control, the $\bar{w}_{\alpha}^i(\cdot)$ were not explicitly introduced in [31], and the upper bounds B_i were handled there by the impulsive control and the L-terms were not needed. Many details are similar to those in §§ 4 and 7. We will outline the steps of the proof.

(1) $B^{\varepsilon}(\cdot)$, $m^{\varepsilon}(\cdot)$ are tight and any weak limit $B(\cdot)$ is continuous. If $\{X^{\varepsilon}(\cdot)\}$ were tight in $D^2[0, \infty)$, then (9.12) holds for any weak limit, all exactly as in theorem 4.6.

(2) $\{\bar{S}_{\alpha}^{i,\varepsilon}(\cdot)\}$ converges weakly to the deterministic function with values $g_{\alpha i} \cdot t$.

(3) The $\{W_{\alpha}^{i,\varepsilon}(\cdot)\}$ are tight owing to the fact that the $\{\delta W_{\alpha,n}^{i,\varepsilon}/\sqrt{\varepsilon}\}$ are martingale differences (with respect to the $\mathcal{F}_{\alpha,n}^{i,\varepsilon}$) and are uniformly square integrable. In addition, the mutual orthogonality of the components for different α, i, imply that the limits are orthogonal continuous martingales with quadratic variation $t/g_{\alpha i}$. The $\bar{W}_{\alpha}^{i,\varepsilon}(\cdot)$ differ from the $W_{\alpha}^{i,\varepsilon}(\cdot)$ only by the scaling $\bar{S}_{\alpha}^{i,\varepsilon}(\cdot)$, which converges as in (2). Thus any weak limit of $\{\bar{W}_{\alpha}^{i,\varepsilon}(\cdot), i = 1, 2, \alpha = a, d\}$ are mutually independent Wiener processes.

(4) Similarly to the case in (3), the $\{M_{\alpha}^{i,\varepsilon}(\cdot)\}$ are also tight and have continuous limits. The limits are orthogonal martingales (with values in R^1 or R^2 according to whether $\alpha = a$ or d, respectively). Since any limits $\bar{M}(\cdot)$ and $M(\cdot)$ are related by the scaling $\bar{S}(\cdot)$, by (2) the quadratic variations satisfy $\langle \bar{M}_{\alpha}^i \rangle(t) = \langle M_{\alpha}^i \rangle(g_{\alpha i} t)$.

(5) $\tilde{Y}^{\varepsilon}(\cdot)$ is also tight and its "limits" are continuous. To see this, we use Theorem 4.4 and prove it for one component only, namely, $\tilde{Y}^{10,\varepsilon}(\cdot)$. Note that

$$Y_n^{10,\varepsilon} - p_{10} Y_n^{1,\varepsilon} = \sqrt{\varepsilon} \sum_1^n \psi_k^1 I_{\{X_k^{1,\varepsilon} = 0\}} (I_k^{10} - p_{10}).$$

The summands are martingale differences (with respect to the intrinsic filtration), due to the fact that $\{I_k^{10,\varepsilon} - p_{10}\}$ are independently and identically distributed and zero mean and independent of $\{\psi_k^1\}$. Thus, for $t = n\varepsilon$,

$$\begin{aligned} &E\{[(Y_{n+m}^{10,\varepsilon} - p_{10} Y_{n+m}^{1,\varepsilon}) - (Y_m^{10,\varepsilon} - p_{10} Y_m^{1,\varepsilon})]^2 \mid Y_k^{1,\varepsilon}, Y_k^{10,\varepsilon}, k \leqq m\} \\ &\quad \leqq (\text{constant}) \cdot \varepsilon[\#k \in (m, m+n] : X_k^{1,\varepsilon} = 0] \\ &\quad \leqq (\text{constant}) \cdot t. \end{aligned} \tag{9.15}$$

This yields the tightness and the continuity of any limit process.

(6) To handle the $Y^\varepsilon(\cdot)$ and $L^\varepsilon(\cdot)$, we use the following slightly modified form of very useful results of Harrison and Reiman [23, Thm. 1]: Recall that the spectral radius of P is less than unity. Let $y(\cdot)$, $f(\cdot)$, and $z(\cdot)$ be in $D^r[0,\infty)$, and *such that* $f(t)=z(t)+(I-P')y(t)$ *and for all* i, $f^i(t)\geqq 0$, $y^i(\cdot)$ *is nonnegative and nondecreasing and increases only when* $f^i(t)=0$. *Then there is a continuous function* (*in the topology of the* sup *norm on bounded time intervals*) $F(\cdot)$ *from* $D^r[0,\infty)$ *to* $D^r[0,\infty)$ *such that* $y(\cdot)=F(z(\cdot))$. *If* $z(\cdot)$ *is continuous, so is* $y(\cdot)$. *There is an analogous result for the equation* $f(t)=z(t)-L(t)$, *where* $f^i(t)$ *is now confined to be* $\leqq B_i$ *where* $B_i>0$.

Using this result in (9.10) yields (loosely speaking), for the time intervals where $X^\varepsilon(t)$ is not in some neighborhood of the corners $(B_1,0)$ or $(0,B_2)$, that $Y^\varepsilon(\cdot)$, $L^\varepsilon(\cdot)$ are continuous and nonanticipative functions of $Z^\varepsilon(\cdot)$, and if the "sections" of $Z^\varepsilon(\cdot)$ on those intervals converge weakly to a continuous process, then so do the corresponding sections of $Y^\varepsilon(\cdot)$ and $L^\varepsilon(\cdot)$.

The corners present no problem (either in our two-dimensional case or in general). To see why, note that, (e.g.), in a neighborhood of $(B_1,0)$, (9.10) reduces to

$$X^{1,\varepsilon}(t)=Z^{1,\varepsilon}(t)-p_{21}Y^{2,\varepsilon}(t)-L^{1,\varepsilon}(t), \qquad X^{2,\varepsilon}(t)=Z^{2,\varepsilon}(t)+(1-p_{22})Y^{2,\varepsilon}(t).$$

Thus, we get $Y^{2,\varepsilon}(\cdot)$ first and then $L^{1,\varepsilon}(\cdot)$ in that neighborhood.

Thus $\{Y^\varepsilon(\cdot), L^\varepsilon(\cdot)\}$ is tight and hence so is $\{X^\varepsilon(\cdot)\}$, and they all have continuous limits, since $\{Z^\varepsilon(\cdot)\}$ does.

(7) Now that we know $\{R^\varepsilon(\cdot)\}$ is tight and the weak limits are continuous, we extract a weakly convergent subsequence and use a "martingale method" analogous to that of Theorem 4.6, but with the dynamical equation (9.10) to show that the limit satisfies (9.11) and that the $\bar M^i_\alpha(\cdot)$ and $w^i_\alpha(\cdot)$ are $\mathcal{R}(t)$-martingales and Wiener processes, respectively. To show that $\tilde Y(\cdot)=$ zero process, we reason as follows. Note that $Y^{i,\varepsilon}(t)=\sqrt{\varepsilon}\,\{\#n: X^{i,\varepsilon}_n=0 \text{ for } n\leqq t/\varepsilon\}$. For given t, the $\{Y^{i,\varepsilon}(t),\varepsilon>0\}$ is bounded in probability by the weak convergence. Thus $\sqrt{\varepsilon}\ Y^{i,\varepsilon}(\cdot)\Rightarrow$zero process. This and (9.15) imply that $\tilde Y^\varepsilon(\cdot)\Rightarrow$zero process.

(8) The quadratic covariation of $\bar M^i_\alpha(\cdot)$ and $\bar w^i_\beta(\cdot)$ with respect to the filtration $\mathcal{R}(t)$ is obtained from the limit as $\varepsilon\to 0$ of the "discrete parameter quadratic covariation":

$$\begin{aligned}(9.16)\qquad &\left\langle \sum_{k=1}^{\varepsilon^{-1}\bar S^{i,\varepsilon}_\alpha(\cdot)} g^{1/2}_{\alpha i}\sigma_{\alpha i}(X^\varepsilon_{S^i_{\alpha,k}})\delta W^i_{\alpha,k+1}, \sum_{k=1}^{\varepsilon^{-1}\bar S^{i,\varepsilon}_\beta(\cdot)} \delta W^i_{\beta,k+1}\right\rangle(t)\\ &\quad=\delta_{\alpha\beta}\delta_{ij}\varepsilon \sum_{k=1}^{\varepsilon^{-1}\bar S^{i,\varepsilon}_\alpha(t)} \sigma_{\alpha i}(X^\varepsilon_{S^i_{\alpha,k}})g^{-1/2}_{\alpha i}.\end{aligned}$$

Choosing a weakly convergent subsequence with limit $x(\cdot)$, $\bar M(\cdot),\cdots$, we get

$$\begin{aligned}(9.17)\qquad \langle \bar M^i_\alpha, \bar w^i_\beta\rangle(t)&=\frac{1}{\sqrt{g_{\alpha i}}}\delta_{ij}\delta_{\alpha\beta}\int_0^{tg_{\alpha i}}\sigma_{\alpha i}\left(x\left(\frac{s}{g_{\alpha i}}\right)\right)ds\\ &=\delta_{ij}\delta_{\alpha\beta}\sqrt{g_{\alpha i}}\int_0^t\sigma_{\alpha i}(x(s))\,ds.\end{aligned}$$

Similarly, we get the quadratic covariation

$$(9.18)\qquad \langle \bar M^i_\alpha, \bar M^i_{\alpha'}\rangle(t)=g_{\alpha i}\int_0^t\sigma_{\alpha i}(x(s))\sigma'_{\alpha i}(x(s))\,ds.$$

The representation (9.14) is implied by these expressions. □

THEOREM 9.2. *Under* (A9.0)-(A9.5), *if* ε *indexes a weakly convergent subsequence of* $\{R^\varepsilon(\cdot)\}$ *with limit* $R(\cdot)$ *as in Theorem* 9.1, *then the costs* $V^\varepsilon(x,m^\varepsilon)$ *converge to*

$V(x, m)$, where $x(\cdot)$ satisfies (9.11) *under $m(\cdot)$ and*

$$V(x,m)=E_x^m\int_0^\infty\int e^{-\beta t}k(x(t),c)m_t(dc)\,dt+E_x^m\int_0^\infty e^{-\beta t}[k_1dL^1(t)+k_2dL^2(t)]. \tag{9.19}$$

Proof. The component of the cost involving $k(\cdot)$ clearly converges, as asserted, by the weak convergence. To get the convergence of the last term in (9.19), we need only show that $\{L^{i,\varepsilon}(n+1)-L^{i,\varepsilon}(n)$, small ε, $i=1,2, n<\infty\}$ is uniformly integrable. We outline part of the calculation for $L^{1,\varepsilon}(\cdot)$. Let $0<\Delta<B_1$. We time the excursions of the process as it goes from the right-hand boundary $x^1=B_1$ to the line $x^1=B_1-\Delta$, and (possibly) back. Fix an integer n. Define

$$\begin{aligned}\tau_0^\varepsilon&=\min\{t\in[n,n+1]: X^{1,\varepsilon}(t)=B_1\},\\ \sigma_{k+1}^\varepsilon&=\min\{t\in[n,n+1]: t\geqq\tau_k^\varepsilon: X^{1,\varepsilon}(t)\leqq B_1-\Delta\},\\ \tau_{k+1}^\varepsilon&=\min\{t\in[n,n+1], t\geqq\sigma_{k+1}^\varepsilon: X^{1,\varepsilon}(t)=B_1\}.\end{aligned}$$

The τ_k^ε and σ_k^ε equal infinity, if not otherwise defined.

For any function $f(\cdot)$, define $\delta_k f=f(\sigma_{k+1}^\varepsilon\wedge(n+1))-f(\tau_k^\varepsilon\wedge n)$. Let N^ε denote the number of k such that $\tau_k^\varepsilon<(n+1)$. Then, by (9.10)

$$\sum_k\delta_kL^{1,\varepsilon}=-\sum_k\delta_kX^{1,\varepsilon}+\sum_k\delta_kZ^{1,\varepsilon}-p_{21}\sum_k\delta_kY^{21,\varepsilon}, \tag{9.20}$$

$$\sum_k\delta_kL^{1,\varepsilon}\leqq(N^\varepsilon+1)B_1+\sum_k\delta_kZ^{1,\varepsilon}. \tag{9.21}$$

Since $\sup_{\varepsilon,n}E|\sum_k\delta_kZ^{i,\varepsilon}|^2<\infty$ and $\sum_k\delta_kL^{1,\varepsilon}=L^{1,\varepsilon}(n+1)-L^{1,\varepsilon}(n)$, we will have the desired result if we prove $\sup_{\varepsilon,n}E(N^\varepsilon)^2<\infty$.

The fact that all moments of the number of excursions from $B_1-\Delta$ to B_1 on the interval $[n, n+1]$ are uniformly bounded in (n,ε) follows from the inequality (9.22), where τ is an arbitrary stopping time in $[n, n+1]$ and $R^\varepsilon(\cdot)$ is defined in Theorem 9.1: There are $\Delta_1>0$ and $\delta_1>0$ such that for all τ, n, and small ε,

$$P\left\{\sup_{s\leqq\Delta_1}|Z^\varepsilon(\tau+s)-Z^\varepsilon(\tau)|\geqq\Delta/2|R^\varepsilon(u), u\leqq\tau\right\}\leqq1-\delta_1\quad\text{w.p.1.} \tag{9.22}$$

(See [38, Thm. 5.3] for more detail on a related calculation.) This ends our discussion of the proof.

Convergence of $V^\varepsilon(x)$ to the minimal cost for the limit problem. Given a filtration $\mathscr{F}_t$ with the $\bar{w}_\beta^\alpha(\cdot)$ being $\mathscr{F}_t$-Wiener processes, the definition of admissible control that we use for (9.11), (9.19) is the same here as in § 3.

To show that $V^\varepsilon(x)=\inf_{m\text{ adm.}}V^\varepsilon(x,m)\to V(x)=\inf_{m\text{ adm.}}V(x,m)$, we use the following additional assumptions. Let $\bar{w}(\cdot)=\{\bar{w}_\alpha^i(\cdot), i=1,2, \alpha=a, d\}$.

(A9.6) For each admissible pair $(m(\cdot), \bar{w}(\cdot))$, there is a unique weak sense solution to (9.11).

(A9.7) For each constant admissible ordinary control, there is a unique strong solution to (9.11), in the sense that $x(t)$ is a measurable function of $\{x=x(0), w(s), m_s(\cdot), s\leqq t\}$.

Remark on (A9.6), (A9.7). The assumption (A9.7) obviously holds in the case of Itô equations under the Lipschitz condition. The assumptions are useful because they guarantee that for each $\rho>0$ there is a ρ-optimal piecewise constant control that depends on the $\bar{w}(\cdot)$ only. Under only weak sense uniqueness, we need to use randomized controls rather than just $\bar{w}(\cdot)$-dependent controls. The approximation theorem is then harder, but can also be carried out. See Chapter 9.3 of [29] for a treatment of the Itô equation case.

We have the following approximation theorems.

THEOREM 9.3. *Assume* (A9.6) *and the continuity of* $b(\cdot,\cdot)$, $\sigma(\cdot)$. *Let* $(m^p(\cdot), \bar{w}(\cdot)) \Rightarrow (m(\cdot), \bar{w}(\cdot))$, *where* $(m^p(\cdot), \bar{w}(\cdot))$ *and* $(m(\cdot), \bar{w}(\cdot))$ *are admissible pairs. Then* (*with* p *indexing the associated processes*) $(x^p(\cdot), Y^p(\cdot), L^p(\cdot), m^p(\cdot), \bar{w}(\cdot)) \Rightarrow (x(\cdot), Y(\cdot), L(\cdot), m(\cdot), \bar{w}(\cdot))$, *satisfying* (9.11). *Also for each* n, $\{L^p(n+1) - L^p(n), n, p < \infty\}$ *is uniformly integrable and* $V(x, m^p) \to V(x, m)$, *if* $k(\cdot,\cdot)$ *is continuous.*

Remark on the proof. The $\{\bar{M}^p(\cdot), m^p(\cdot), B^p(\cdot), \bar{w}(\cdot)\}$ is obviously tight and has continuous limits. Then, as in Theorem 9.1, the same result holds for $\{L^p(\cdot), Y^p(\cdot), x^p(\cdot)\}$. The limit of any weakly convergent subsequence must satisfy (9.11) for the given $(m(\cdot), \bar{w}(\cdot))$. By uniqueness, the entire sequence converges as stated. The uniform integrability of $\{L^p(n+1) - L^p(n), n, p < \infty\}$ is proved as in Theorem 9.2. The convergence $V(x, m^p) \to V(x, m)$ follows from this and the weak convergence.

THEOREM 9.4. *Under* (A9.7) *and the assumptions of Theorem* 9.3, *for each* $\rho > 0$ *there is a* $\delta > 0$ *and a* ρ*-optimal ordinary admissible piecewise constant control* (*constant on the intervals* $[i\delta, i\delta + \delta)$, $i = 0, 1, \cdots$) *and taking only finitely many values. Furthermore, the control used on the interval* $[k\delta, k\delta + \delta)$ *can be taken to be a function of* $\{\bar{w}(j\theta), j\theta \leqq k\delta\}$ *for small enough* θ, *and where the "boundaries of the decision sets" have zero probability* (*as in Theorem* 7.1 *or* 8.3).

Remark on the proof. We first use Theorem 9.3 to get a piecewise constant $\rho/2$-optimal control. We then use (A9.7) and an argument of the sort used in Theorem 7.1 to show that on $[k\delta, k\delta + \delta)$, the control can be taken to be a function of $\{\bar{w}(s), s \leqq k\delta\}$. Then, a further approximation, as in Theorem 7.1, is used to show that we can use the "samples" for small enough θ.

Finally, we can state the following theorem.

THEOREM 9.5. *Under* (A9.0)-(A9.7), $V^\varepsilon(x) \to V(x)$.

Remark on the proof. The proof is very similar to those of Theorems 7.1 or 8.3. By Theorem 9.2 and the definition of $V(x)$, we have $\lim_\varepsilon V^\varepsilon(x) \geqq V(x)$. To get the reverse inequality, we apply the ρ-optimal control described in Theorem 9.4. Let $F_k(\bar{w}(j\theta), j\theta \leqq k\delta)$ denote the control used on $[k\delta, k\delta + \delta)$. Then for controlling $X^\varepsilon(\cdot)$ on the time interval $[k\delta, k\delta + \delta)$, use $F_k(\bar{w}^{i,\varepsilon}_\alpha(j\theta), j\theta \leqq k\delta, i = 1, 2, \alpha = a, d)$. The proof concludes via a weak convergence argument (as in Theorem 7.1) to show that the limit process is precisely the one (9.11) associated with the ρ-optimal decision functions $\{F_k\}$.

9.1. The numerical method.

The approximating Markov chain. The controlled Markov chain approximation $\{\xi^h_n\}$ is constructed as in the previous sections, the only variation being that we account for the boundary reflection in a consistent way. For simplicity of presentation we use a combination of the methods of §§ 5.1 and 5.2 for the construction of the approximating chain. Define G_h to be the h-grid on $G = [0, B_1] \times [0, B_2]$ and let the B_i be integral multiples of h. Let G^+_h denote the h-gridpoints on $[-h, B_1 + h] \times [-h, B_2 + h]$. To account for the reflection, it is convenient to compute the transition probability for $x \in G_h$ in two steps: first compute the probabilities as if there were no reflection; then, if the transition were actually to a point $y \in G^+_h - G_h$, to immediately reflect back to G_h in a way that is consistent with the behavior of (9.11) on the boundary. The actual state space is G^+_h. But, since we use $\Delta t^h(x, c) = 0$ for $x \in G^+_h - G_h$, we have instantaneous reflection.

Part 1. Let $x \in G_h$. Ignoring the boundary reflection terms, (9.11) reduces to

$$dx = \begin{pmatrix} d\bar{M}_{a1} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ d\bar{M}_{a2} \end{pmatrix} + d\bar{M}_{d1} + d\bar{M}_{d2} + dt \int b(x, c) m_t(dc). \tag{9.23}$$

Let $p^h(x, y \mid c)$ denote the consistent Markov chain transition probabilities for model (9.23), and $\Delta t^h(x, c)$ the associated interpolation interval. These satisfy (2.4) and can be obtained by the methods of § 5, since (9.23) is just an Itô equation.

Part 2. Let $x \in G_h^+ - G_h$. Write $x = (x^1, x^2)'$. (a) Suppose that some $x^i > B_i$, but no $x^i < 0$. Then simply reflect back instantaneously to the nearest point on G_h.

(b) Let some $x^i < 0$, but no $x^i > B_i$. Then find a vector $\delta \bar{Y}^h = (\delta \bar{Y}^{1,h}, \delta \bar{Y}^{2,h})'$ where $\delta \bar{Y}^{i,h} > 0$ only if $x^i < 0$ (and is zero otherwise) and such that $x + (I - P')\delta \bar{Y}^h = \tilde{x}$ is on the boundary of the rectangle G. Such a $\delta \bar{Y}^h$ can be found since the spectral radius of P is less than unity. Generally, $\tilde{x}$ will not be a gridpoint. We then "randomize" the transitions, so that the mean value remains $\tilde{x}$. For example, refer to Fig. 9.2. There $(1 - p_{11})\delta \bar{Y}^1 = h$ and $\delta \bar{Y}^2 = 0$ so that $q_1 = p_{12}/(1 - p_{11})$ (q_1 is shown in the figure). Let x' and x'' denote the nearest gridpoints. Then use the (actually uncontrolled) transition probabilities $p^h(x, x' \mid c) = q_2$, $p^h(x, x'' \mid c) = q_1 = 1 - q_2$. In the limit, as $h \to 0$, the randomized transition has the same effect as the "mean" transition to $\tilde{x}$, by the same logic that allowed us to replace, (e.g.), Y^{i0} by $p_{i0} Y^i$ in Theorem 9.1.

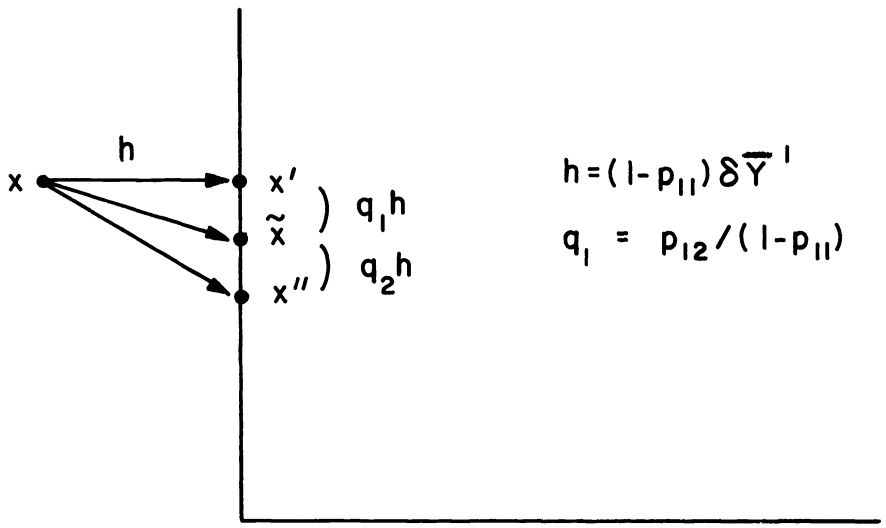


FIG. 9.2. *Example of the reflection model.*

(c) Let some $x^i > B_i$ and some $x^i < 0$. Here, the choices of $\delta \bar{Y}^h$ and δL^h "decouple" as illustrated in part (6) of the proof of Theorem 9.1. We first calculate $\delta \bar{Y}^h$ and randomize as in (b), then take up any "residue" by repeating the procedure of (a).

The dynamic programming equation. The proper transition probabilities for $\{\xi_n^h\}$ are given by a concatenation of those in Parts 1 and 2. With this concatenation, the chain cannot leave G_h. For the numerical calculations, however, it is not necessary to combine the two steps, and the procedure that we use is for the states space G_h^+ with the transition probabilities calculated on G_h and $G_h^+ - G_h$, as above.

For $x \in G$, under control value c the incremental cost is $\Delta t^h(x, c)k(x, c)$ and the discount factor is $\exp - \beta \Delta t^h(x, c)$. For $x \notin G$, we use $\Delta t^h(x, c) = 0$, and the appropriate cost is that indicated in (9.25) below. If $x^1 = B_1 + h$ and $x^2 \geqq 0$, then we must reduce x^1 by h, with a cost $k_1 h$—to be consistent with (9.19). Suppose that $x^1 = B_1 + h$ and $x^2 = -h$. Then, of course, we must reflect back to the set G_h, as in (c) above. Part of the "mean reflection" is taken up by the $\delta \bar{Y}^{2,h}$ term calculated above. The mean value

of $\delta L^{1,h}$ is $h(1-p_{21}/(1-p_{22}))$, analogously to the calculation used for Fig. 9.2, and this explains the appearance of this term in (9.25) below. With $V^h(x)$ used to denote the optimal cost for the Markov chain, the dynamic programming equation is, for $x \in G_h$,

$$V^h(x) = \min_{c \in U} \left[e^{-\beta \Delta x^h(x,c)} \sum_y p^h(x, y \mid c) V^h(y) + \Delta t^h(x, c) k(x, c) \right]. \tag{9.24}$$

For $x \in G_h^+ - G_h$, $\Delta t^h(x, c) = 0$, the transition probabilities (actually uncontrolled) are given in Part 2 above, and

$$\begin{aligned} V^h(x) = \Bigg[& \sum_y p^h(x, y \mid c) V^h(y) + k_1 h I_{\{x^1 > B_1, x^2 \geqq 0\}} \\ & + k_1 h \left(1 - \frac{p_{21}}{1-p_{22}}\right) I_{\{x^1 > B_1, x^2 < 0\}} + k_2 h I_{\{x^2 > B_2, x^1 \geqq 0\}} \\ & + k_2 h \left(1 - \frac{p_{12}}{1-p_{11}}\right) I_{\{x^2 > B_2, x^1 < 0\}} \Bigg]. \end{aligned} \tag{9.25}$$

Equations (9.24) and (9.25) can be solved by any of the usual methods for the discounted problem.

On the convergence $V^h(x) \to V(x)$. We can state the following theorem.

THEOREM 9.6. *Let $b(\cdot, \cdot)$ and $\sigma(\cdot) = (\sigma_{\alpha i}, \alpha = a, d, i = 1, 2)$ be continuous, assume* (A9.4)-(A9.7), *and let each U_α be compact. Let the spectral radius of P be less than unity. Then, for the chain $\{\xi_n^h\}$ constructed above in this section, $V^h(x) \to V(x)$.*

Remark. The proof is quite similar to that given in § 7, but with the technique of this section used to treat the reflection terms. Since we are approximating the limit problems (9.11), (9.19) directly, no heavy traffic analysis or assumptions are needed.

10. Controlled reflected diffusions. There are two convenient models to use for the reflected diffusion process, the submartingale formulation of Stroock and Varadhan [47], [34], to be discussed briefly at the end of the section, and the Skorokhod problem formulation to which we now turn.

We will use the following assumption:

(A10.1) G is the closure of a bounded open set with a twice continuously differentiable boundary. Let $n(x)$ denote the outward normal to ∂G at x, and let $\gamma(x)$ denote the reflection direction. Suppose that $\gamma(\cdot)$ is the restriction to ∂G of a function that is twice continuously differentiable in a neighborhood of ∂G and let there be $\alpha_0 > 0$ such that $-\gamma'(x) n(x) \geqq \alpha_0$, for all $x \in \partial G$.

THE SKOROKHOD PROBLEM. Let $\{\mathcal{F}_t\}$ be a filtration on some probability space and let $w(\cdot)$ be an $\mathcal{F}_t$-standard Wiener process. We say that $x(\cdot)$ solves the (uncontrolled) Skorokhod problem if it is $\mathcal{F}_t$-adapted, continuous, and there is continuous $\mathcal{F}_t$-adapted $Y(\cdot)$ such that for $x \in G$ (var denotes variation),

$$\begin{aligned} & x(t) = x + \int_0^t b(x(s))\, ds + \int_0^t \sigma(x(s))\, dw(s) + Y(t), \\ & (\text{var } Y)(t) \equiv |Y|(t) = \int_0^t I_{\{x(s) \in \partial G\}}\, d|Y|(s), \\ & Y(t) = \int_0^t \gamma(x(s))\, d|Y|(s), \; x(t) \in G. \end{aligned} \tag{10.1}$$

The controlled reflected diffusion. An admissible pair $(w(\cdot), m(\cdot))$ is defined as in § 3. The associated reflected diffusion model is (10.1) but with $m(\cdot)$ added, namely,

$$x(t) = x + \int_0^t \int b(x(s), c) m_s(dc)\, ds + \int_0^t \sigma(x(s))\, dw(s) + Y(t). \tag{10.2}$$

The solution to (10.2) is said to be a *strong solution* if for each t, $x(t)$, $Y(t)$ are measurable on the σ-algebra induced by $\{w(s), m(s, \cdot), s \leqq t\}$. The basic results we use are in [40]. Reference [40] does not explicitly use time or control dependence. But, under the Lipschitz and continuity condition in (A2.1) all the derivations and results hold when the control is added. A basic result is the following "controlled" version of Theorem 4.3 of [40].

THEOREM 10.1. *Assume* (A2.1) *and* (A10.1), *with* $(m(\cdot), w(\cdot))$ *being an admissible pair. Then there is a unique strong solution* $x(\cdot)$, $Y(\cdot)$ *to* (10.2).

The boundary condition (A10.1) can be weakened [40, Thm. 4.4] to include piecewise smooth boundaries with "convex corners" and possibly multivalued reflection directions at the corners—but not including all cases dealt with in § 9. The particular restrictions on ∂G and on the consequent chains allow a relatively simple discussion of the general idea. The method is extended below. The technique of the proof of Theorem 4.3 in [40] yields the following theorem.

THEOREM 10.2. *Assume* (A2.1) *and* (A10.1). *Let* $(m^n(\cdot), w^n(\cdot))$ *be an admissible pair for each* n, *with* $(m^n(\cdot), w^n(\cdot)) \Rightarrow (m(\cdot), w(\cdot))$. *Then there is a filtration* $\mathcal{F}_t$ *such that* $w(\cdot)$ *is a standard* $\mathcal{F}_t$*-Wiener process and* $(m(\cdot), w(\cdot))$ *is an admissible pair. If* $(x^n(\cdot), Y^n(\cdot))$ *solve* (10.2) *with* $(m^n(\cdot), w^n(\cdot))$ *used, then* $(x^n(\cdot), Y^n(\cdot), m^n(\cdot), w^n(\cdot)) \Rightarrow (x(\cdot), Y(\cdot), m(\cdot), w(\cdot))$, *satisfying* (10.2).

In order to get the tightness for the sequence of interpolated Markov chains we need the following theorem.

THEOREM 10.3 [40, Thm. 4.1]. *Assume* (A10.1) *and consider the Skorokhod problem*:

$$x(t) = f(t) + k(t), \tag{10.3}$$

where $f(\cdot)$ *and* $k(\cdot)$ *are in* $C^r[0, T]$ (*the space of* R^r*-valued continuous functions on* $[0, T]$), $f(0) \in G$, $k(\cdot)$ *is of bounded variation on* $[0, T]$ *and*

$$k(t) = \int_0^t \gamma(x(s)) d|k|(s), \qquad |k|(t) = \int_0^t I_{\{x(s) \in \partial G\}} d|k|(s).$$

If $f(\cdot)$ *is in a compact set in* $C^r[0, T]$, *then* $(x(\cdot), k(\cdot), |k|(\cdot))$ *are in a compact set in* $C^{2r+1}[0, T]$.

The cost functions. As in the preceding sections, we use the discounted cost

$$\begin{aligned} V(x, m) &= E_x^m \int_0^\infty e^{-\beta t} \int k(x(s), c) m_t(dc)\, dt, \\ V(x) &= \inf_{m \text{ adm.}} V(x, m), \end{aligned} \tag{10.4}$$

for illustrative purposes. All the usual forms of the cost function can be used. Also the singular and impulsive control problems can be treated.

The Markov chain approximation. As in the previous sections, any chain that is consistent with (2.4) in G and with the boundary reflection direction $\gamma(x)$, if the process "attempts to leave G," will work. For the sake of simplicity of exposition we illustrate *one* procedure for a two-dimensional problem. But it should be clear that there are many variations of the method that will work in a space of any finite dimension.

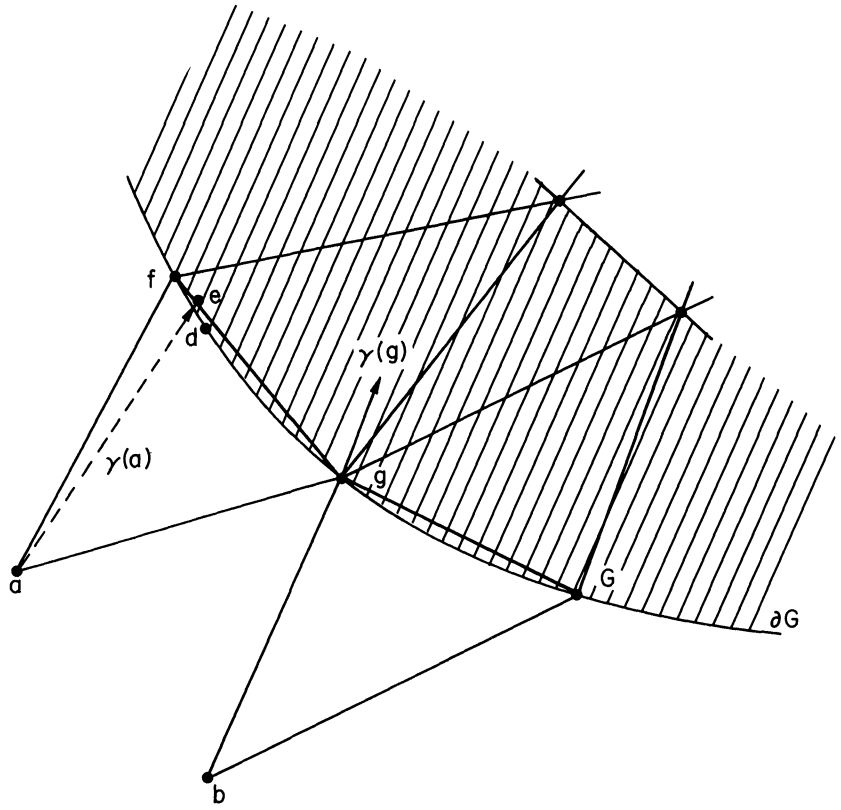


FIG. 10.1. *The triangulation*: *edges* $= O(h)$.

We let G be "triangulated," following the scheme outlined in § 5.2. The sides of the triangles are $O(h)$. Of course, the triangulation might be based on a finite-difference grid. We suppose the situation illustrated in Fig. 10.1, where the lines connecting gridpoints in G do not leave G and the only lines leaving G start on ∂G. As in Fig. 10.1, we extend the triangulation outside G by including all points reachable along the sides of the triangles emanating from points in G. Let the gridpoints not in G be denoted by ∂G_h^+; the gridpoints in G are denoted by G_h.

For $x \in G_h$, let $p^h(x, y \mid c)$ denote any continuous transition probability satisfying (2.4), with continuous interpolation intervals $\Delta t^h(x, c) \geqq k_1, h^2 > 0$. Now, let $x \in \partial G_h^+$, and let y denote the points in G_h connected to x by an edge. Let $p^h(x, y)$ (not depending on c) be transition probabilities such that $(\Sigma' y\, p^h(x, y) - x) \equiv \delta \hat{Y}(x)$ points in the "reflection direction" $\gamma(x)$. The interpolation times are $\Delta t^h(x) = 0$ for such states. Thus, they are "instantaneous." A more concrete construction is given below.

The dynamic programming equation. As in § 9, the dynamic programming equation is

$$
\begin{aligned}
V^h(x) &= \min_{c \in U} \left[e^{-\beta \Delta t^h(x,c)} \sum_y p^h(x, y \mid c) V^h(y) + \Delta t^h(x, c) k(x, c) \right], \qquad x \in G_h, \\
V^h(x) &= \sum_y p^h(x, y) V^h(y), \qquad x \in \partial G_h^+.
\end{aligned} \tag{10.5}
$$

A construction of $p^h(x, y)$ for $x \in \partial G_h^+$. Refer to Fig. 10.1. Let $x = \xi_n^h = (a)$. Draw the line from x in direction $\gamma(x)$, hitting the first edge in G at (e). "Randomize" (e) by writing it as a convex combination of (f) and (g): $(e) = p^h(a, f)(f) + p^h(a, g)(g)$. This technique yields the desired $p^h(x, y)$, $x \in G_h^+$, in general.

Notation. Let $\{\xi_n^h\}$ denote the constructed chain, but *where the instantaneous states are ignored.* Thus if $\xi_n^h = (g) \to (a) \to (f)$, we have $\xi_{n+1}^h = (f)$, and let $\xi^h(\cdot)$ be the continuous parameter interpolation (intervals Δt_n^h, as always). Let $\xi_{n+1}^{h,1}$ denote the successor state to ξ_n^h, whether instantaneous or not. Thus if $\xi_{n+1}^{h,1} \in G_h$, $\xi_{n+1}^{h,1} = \xi_{n+1}^h$. For $\xi_{n+1}^{h,1}$ instantaneous, let $\delta \hat{Y}_n^h \equiv \delta \hat{Y}(\xi_{n+1}^{h,1})$ denote the vector that takes $\xi_{n+1}^{h,1}$ to the nearest edge in G in direction $\gamma(\xi_{n+1}^{h,1})$; e.g., (e)-(a), if $\xi_{n+1}^{h,1} = (a)$ in the figure ($\delta \hat{Y}_n^h = 0$ if $\xi_{n+1}^{h,1} \in G$). Let $\xi_{n+1}^{h,2} = \xi_{n+1}^{h,1} + \delta \hat{Y}_n^h$, for $\xi_{n+1}^{h,1}$ instantaneous. Finally, define $\delta \tilde{Y}_n^h = \xi_{n+1}^h - \xi_{n+1}^{h,2}$, for $\xi_{n+1}^{h,1}$ instantaneous, with $\delta \tilde{Y}_n^h = 0$ otherwise. Thus $\delta \tilde{Y}_n^h = (f)$-$(e)$ or (g)-(e) if $\xi_{n+1}^{h,2} = (e)$. Define $\hat{Y}^h(t) = \sum_{t_{n+1}^h \le t} \delta \hat{Y}_n^h$.

The dynamical equations. We can now write the dynamical equations in a way that will be convenient for relating $\xi^h(\cdot)$ to the solution of the Skorokhod problem. Define $\beta_n^h = (\xi_{n+1}^{h,1} - \xi_n^h) - E_n^h(\xi_{n+1}^{h,1} - \xi_n^h)$, similarly to the definition above (4.9). Then write

$$\xi_{n+1}^h = \xi_n^h + [\xi_{n+1}^{h,1} - \xi_n^h] + [\delta \hat{Y}_n^h + \delta \tilde{Y}_n^h] I_{\{\xi_{n+1}^{h,1} \notin G\}}, \tag{10.6}$$

$$\xi_{n+1}^{h,1} = \xi_n^h + b(\xi_n^h, u_n^h)\Delta t_n^h + \beta_n^h + O(h^\alpha \Delta t_n^h) + [\delta \hat{Y}_n^h + \delta \tilde{Y}_n^h]. \tag{10.7}$$

THEOREM 10.4. *Assume* (A2.1), (A2.2), (A10.1), *and* $\inf_{x,c} \Delta t^h(x, c) \ge k_1 h^2$ *for some* $k_1 > 0$, *and assume the Markov chain approximation selected above. Let* $\{u_n^h\}$ *be any admissible control for* $\{\xi_n^h\}$, *and let* m_n^h *denote its relaxed control representation. Define* $m^h(\cdot)$ *by its derivative:* $m_t^h = m_n^h$ *on* $[t_n^h, t_{n+1}^h)$. *Then* $\{\xi^h(\cdot), m^h(\cdot), \hat{Y}^h(\cdot)\}$ *is tight. If* $\{(x(\cdot), m(\cdot), Y(\cdot))\}$ *is the limit of a weakly convergent subsequence, then there is a* $w(\cdot)$ *such that* $(m(\cdot), w(\cdot))$ *is an admissible pair and* (10.2) *holds. Also* $V^h(x) \to V(x)$.

Proof. The proof follows the lines of § 7, suitably modified to account for the reflection, and we only give the details of the parts that differ significantly from the previous proofs. Define

$$\beta^h(t) = \sum_{t_{n+1}^h \le t} \beta_n^h, \qquad B^h(t) = \sum_{t_{n+1}^h \le t} b(\xi_n^h, u_n^h)\Delta t_n^h.$$

Let us relate (10.7) to the Skorokhod problem. For any piecewise constant function, let the overbar denote the piecewise linear interpolation. Let $t \in [t_n^h, t_{n+1}^h)$ with $\xi_{n+1}^{h,1} \in \partial G_h^+$ and write ($\varepsilon_n^h = O(h^\alpha \Delta t_n^h)$)

$$\bar{\xi}^h(t) = \xi_n^h + [b(\xi_n^h, u_n^h)\Delta t_n^h + \beta_n^h + \varepsilon_n^h + \delta \tilde{Y}_n^h](t - t_n^h)/\Delta t_n^h + \delta \hat{Y}_n^h (t - t_n^h)/\Delta t_n^h. \tag{10.8}$$

Due to the curvature of the boundary, $\bar{\xi}^h(\cdot)$ is not necessarily on ∂G (it is the "secant line" in Fig. 10.1), but it is within $O(h^2)$ of ∂G. Add this "error" $\eta^h(t)$ to both sides of (10.8). Then, for $t \in [t_n^h, t_{n+1}^h)$,

$$\begin{aligned} \bar{\xi}^h(t) + \eta^h(t) = &\left\{ [\xi_n^h + b(\xi_n^h, u_n^h)\Delta t_n^h + \beta_n^h + \varepsilon_n^h + \delta \tilde{Y}_n^h] \frac{(t - t_n^h)}{\Delta t_n^h} + \eta^h(t) \right\} \\ &+ \delta \hat{Y}_n^h \frac{(t - t_n^h)}{\Delta t_n^h}. \end{aligned} \tag{10.9}$$

Now, relate (10.9) to the Skorokhod problem, where the term in braces is the $f(\cdot)$ in (10.3). The $\delta \hat{Y}_n^h (t - t_n^h)/\Delta t_n^h$ is not quite the reflection term $k(\cdot)$ in (10.3) on the interval $[t_n^h, t_{n+1}^h)$, due to the curvature of ∂G, but it is within $O(h^2)$ of the correct reflection term. Define $Y^h(\cdot)$ by letting $Y^h(y) - Y^h(t_n^h) = \delta \hat{Y}_n^h (t - t_n^h)/\Delta t_n^h + O(h^2)$ be the correct

reflection term in $[t_n^h, t_{n+1}^h)$. Then, we can write for all $t \geqq 0$,

$$\bar{\xi}^h(t) + \eta^h(t) = Z^h(t) + Y^h(t), \tag{10.10}$$

where

$$Z^h(t) = x + \bar{B}^h(t) + \bar{\beta}^h(t) + \bar{\varepsilon}^h(t) + \bar{\tilde{Y}}^h(t) + \mu^h(t),$$

where $\mu^h(t)$ collects the $O(h^2)$ terms. In particular,

$$\mu^h(t) = O(h^2)[\#n : t_n^h \leqq t \text{ and } \xi_{n+1}^{h,1} \in G_h^+], \tag{10.11}$$

$$Y^h(t) = O(h)[\#n : t_n^h \leqq t \text{ and } \xi_{n+1}^{h,1} \in G_h^+]. \tag{10.12}$$

Also, $\eta^h(\cdot)$ and $Y^h(\cdot) - \hat{Y}^h(\cdot)$ are of the same order as $\mu^h(\cdot)$.

As in § 7, $\{B^h(\cdot), \beta^h(\cdot), \varepsilon^h(\cdot)\}$ is tight and the limits are continuous, hence also for the $\bar{B}^h(\cdot), \cdots$. Also $\bar{\varepsilon}^h(\cdot) \Rightarrow$ zero process. By (10.11) and the fact that $\Delta t_n^h \geqq h^2 k_1$, $\{\mu^h(\cdot)\}$ is also tight and has continuous limits. We have

$$\begin{aligned} &E[\delta\tilde{Y}_{n+1}^h \mid \delta\tilde{Y}_i^h, i \leqq n] = 0, \qquad \delta\tilde{Y}_n^h = O(h), \\ &E[|\delta\tilde{Y}_{n+1}^h|^2 \mid \delta\tilde{Y}_i^h, i \leqq n] = O(h^2). \end{aligned} \tag{10.13}$$

Equations (10.13) imply the tightness of $\{\tilde{Y}^h(\cdot)\}$ (hence of $\{\bar{Y}^h(\cdot)\}$) and the continuity of the limits. Thus $\{Z^h(\cdot)\}$ is tight and has continuous limits.

By Theorem 10.3, $\{\xi^h(\cdot) + \eta^h(\cdot), Y^h(\cdot)\}$ is tight and has continuous limits. This implies that $\{Y^h(t)\}$ is bounded in probability. By this and (10.11) and (10.12), we have that $\mu^h(\cdot) \Rightarrow$ zero process, and similarly $\eta^h \Rightarrow$ zero process. Thus, (10.10) can be rewritten as

$$\xi^h(t) = x + B^h(t) + \beta^h(t) + Y^h(t) + \text{“small” error}, \tag{10.14}$$

where all the functions in (10.14) are tight and have continuous limits. From this point on, the proof is almost the same as in § 7, and the details are omitted. □

An alternative approach. The framework above is useful since uniqueness of the solution to (10.2) is known by Theorem 10.1, and Theorem 10.3 yields tightness of $\{Y^h(\cdot)\}$. The condition (A10.1) can be weakened, at the expense of assuming uniqueness. The following method follows a suggestion made to the author by Michael Taksar, who is using related ideas for the study of the Skorokhod problem. Let G have a continuous boundary and define G, G_h, and ∂G_h^+ as before. Let the sides of the edges be proportional to h. Let $\gamma(\cdot)$ be continuous with $\inf_{x \in \partial G} |\gamma(x)| > 0$, and suppose that there are $r > 0$, $\rho > 0$, such that the cone with vertex x, radius r, and centerline $x + \gamma(x)\rho$ is in G for all $x \in \partial G$. For $x \in G_h$, choose $p^h(x, y \mid c)$ as before. For $x \in \partial G_h^+$, choose (uncontrolled) $p^h(x, y)$ such that $\sum p^h(x, y)y - x = \gamma(x)h + o(h)$. The $x \in \partial G_h^+$ comunicate with $y \in G_h$ as before, but these y need not be on ∂G. Suppose that for each $\delta > 0$, there is a δ-optimal admissible pair $(m(\cdot), w(\cdot))$ such that the Skorokhod problem (10.2) has a unique solution. Let $b(\cdot)$, $\sigma(\cdot)$, and $k(\cdot)$ be bounded and continuous. Then $V^h(x) \to V(x)$.

Except for the fact that Theorem 10.3 can no longer be used to get tightness of $\{Y^h(\cdot)\}$ and that uniqueness must be assumed, the proof is essentially the same as that of Theorem 10.4. Tightness of $\{Y^h(\cdot)\}$ follows from the fact that if the set were not tight, then the “cone” condition would imply that the “reflection” terms push the $\xi^h(\cdot)$ “far” away from ∂G, which is a contradiction to the fact that the reflection terms only act at states in ∂G_h^+.

Other approximations. The above described numerical methods are reasonably easy to use, and are similar to any reasonable alternative—since the appropriate consistency is required for any one. Given small $\delta > 0$, we can approximate the reflected diffusion by a process which, when hitting ∂G at a point x, jumps a distance δ in direction $\gamma(x)$. As $\delta \to 0$, the optimal cost $V_\delta(x)$ for this problem converges to $V(x)$. Thus, we can use a numerical procedure for the optimal control problem for the altered process for small δ.

The submartingale problem formulation. The martingale problem representation [49] which was used in Theorem 4.6 is a very convenient way of characterizing a diffusion process, and, in particular, of showing that the limit of a weakly convergent sequence of processes is a diffusion process. There is an analogous characterization of reflected diffusions which is, in some ways, more general than the Skorokhod problem approach—in that it allows for "sticky" boundaries or "delayed reflection" (but it does not allow for flows on boundaries). The basic paper is [47]. An approach to using it for the uncontrolled numerical problem is in [34], but the addition of controls is similar to what was done in this paper. Only a brief description will be given. We assume the following.

(A10.2) There is $\alpha_1 > 0$ such that $n'(x)a(x)n(x) \geqq \alpha_1 > 0$, for all $x \in \partial G$.

Let G, $\gamma(\cdot)$ be as in (A10.1) and let $\rho \geqq 0$. We say that a continuous process $x(\cdot)$ solves the submartingale problem for operator $\mathscr{L}$ and boundary reflection $\gamma(\cdot)$ and "stickiness" ρ if for each smooth $f(\cdot)$ with compact support and satisfying $\rho F_t(x, t) + \gamma'(x)F_x(x) \geqq 0$, the process

$$S_f(t) = f(x(t), t) - f(x, 0) - \int_0^t [f_s(x(s), s) + \mathscr{L}f(x(s), s)] I_{\{x(s) \in G^0\}}\, ds \tag{10.15}$$

is a submartingale. Then [47] there is a nondecreasing continuous process $\mu(\cdot)$ and a standard Wiener process $w(\cdot)$ such that $x(\cdot)$, $\mu(\cdot)$ are nonanticipative with respect to $w(\cdot)$ and

$$\begin{aligned} x(t) = x &+ \int_0^t b(x(s)) I_{\{x(s) \in G^0\}}\, ds + \int_0^t \sigma(x(s)) I_{\{x(s) \in G^0\}}\, dw(s) \\ &+ \int_0^t \gamma(x(s)) I_{\{x(s) \in \partial G\}}\, d\mu(s). \end{aligned} \tag{10.16}$$

If $\rho = 0$, then $\mu(\cdot)$ is singular with respect to Lebesgue measure, the total time spent on ∂G is zero, and the submartingale problem and the Skorokhod problem have the same solution, under compatible conditions. If $\rho > 0$, then $\mu(\cdot)$ is absolutely continuous with respect to Lebesgue measure, and the time spent on the boundary might not be zero.

We can easily define a control problem and we can, if desired, have different costs on ∂G and on G^0. In [34], it is proved that the time that the sequence $\xi^h(\cdot)$ of interpolated Markov chains spends in the set $N_\varepsilon(\partial G) - \partial G$ on any interval $[0, T]$ goes to zero in the mean as $\varepsilon \to 0$, uniformly in (small) h. This fact allows us to consider separate costs for ∂G and G^0 and still get the desired weak convergence. The Markov chain used in [34] can be replaced by any one satisfying the appropriate consistency conditions.

11. The average cost per unit time problem. In this section, the basic system equation will be either (9.11) where Y and L satisfy the conditions of Theorem 9.1 (where B^i and $\bar{M}^i$ are given by (9.12) and (9.14), respectively) or (10.2), under (A10.1).

We use the average cost per unit time:

$$\gamma(m) = \overline{\lim_t} \frac{1}{t} \int_0^t \int E_x^m k(x(s), c) m_s(dc)\, ds, \qquad \bar{\gamma} \equiv \inf \gamma(m), \tag{11.1}$$

for admissible $m(\cdot)$. In order to do the numerical problem, we need to work in a bounded region. For specificity, we adopt the reflected diffusion model used in either § 9 or 10 via (A11.1).

(A11.1) G and the boundary reflection directions satisfy the conditions of the first sentence of this section. The approximating Markov chain $\{\xi_n^h\}$ has a single recurrent class under each feedback control. Let $\Delta t^h(x, c)$ not depend on c. There is $q_0 > 0$ such that $\inf_x \Delta t^h(x) \geqq q_0 h^2$.

The "single recurrent chain" condition is not necessary—it simply saves some additional detail in the development.

We are concerned with the average cost per unit time problem for $x(\cdot)$ (or for its approximation $\xi^h(\cdot)$), but not directly for the chain $\{\xi_n^h\}$. If $\Delta t^h(x)$ did not depend on x, then we could approximate the average cost per unit time for $x(\cdot)$ by that for $\{\xi_n^h\}$. But, if $\Delta t^h(x)$ is x-dependent, we need to weigh the values obtained for $\{\xi_n^h\}$ according to the occupancy times used in the interpolation. Of course, as discussed in § 2, there are numerical advantages to using the appropriate x-dependence in $\Delta t^h(x)$.

The dynamic programming equation for the Markov chain. We now show how to get the appropriate approximation. We start by proceeding formally and ignoring the boundary ∂G. If $\bar{\gamma}$ is the optimal cost then, under appropriate conditions [5], [10], [36], there is a smooth function $V(\cdot)$ such that $(\bar{\gamma}, V(\cdot))$ satisfy

$$\bar{\gamma} = \min_{c \in U} [\mathscr{L}^c V(x) + k(x, c)]. \tag{11.2}$$

Conversely, any solution $(\gamma, V(\cdot))$ to (11.2) implies that $\gamma = \bar{\gamma}$, if $E^u V(x(t))/t \to 0$ under the minimizing $u(\cdot)$, as $t \to \infty$. See [33] for some related formal calculations.

In order to get the appropriate dynamic programming equation for the discrete problem, let us apply the finite-difference approximations of § 5.1 to (11.2). (This will be generalized below.) Letting γ^h, $V^h(\cdot)$, denote the finite difference solution on a grid G_h, we get

$$V^h(x) = \min_{c \in U} \left[\sum_y p^h(x, y \mid c) V^h(y) + \Delta t^h(x)(k(x, c) - \gamma^h) \right], \qquad x \in G_h. \tag{11.3}$$

The $p^h(x, y \mid c)$ are those from § 5.1. This is, in fact, a dynamic programming equation for a semi-Markov decision process. Equation (11.3) can be solved by the approximation in policy space method [46], [52].

Now we reintroduce the boundary, under either set of conditions in (A11.1). For $x \in G_h$, use (11.3). For $x \in G_h^+$ use (instantaneous reflection, $\Delta t^h(x) = 0$ for $x \in \partial G_h^+$)

$$V^h(x) = \sum_y p^h(x, y \mid c) V^h(y), \tag{11.4}$$

where the $p^h(x, y \mid c)$ are those used in (§ 9, part 2) or in the second line of (10.5), according to the case. Now, we drop the specificity of the above $p^h(x, y \mid c)$ for $x \in G_h$, and use any of the transition functions that can be used in §§ 9 or 10, according to the case. Equations (11.3) and (11.4) are the correct dynamic programming equations for our approximation—they appropriately incorporate the holding times $\Delta t^h(x)$, as will be seen below.

The cost function for a fixed control. For a continuous feedback control $u(\cdot)$, let $V^h(\cdot, u)$ and $\gamma^h(u)$ denote the solution to the equation

$$\begin{aligned} V^h(x,u) &= \sum_y p^h(x,y \mid u(x)) V^h(y,u) + (k(x,u(x)) - \gamma^h(u))\Delta t^h(x), \qquad x \in G_h, \\ V^h(x,u) &= \sum_y p^h(x,y \mid u(x)) V^h(y,u), \qquad x \in \partial G_h^+. \end{aligned} \tag{11.5}$$

Under the uniqueness of the recurrent class in (A11.1), (11.5) has a unique solution $\gamma^h(u)$. The solution $V^h(\cdot, u)$ is not unique, since for any constant K, $V^h(\cdot, u)+K$ is also a solution. But, if we restrict $V^h(\cdot, u)$ such that $V^h(x_0, u)=0$ for some x_0, then the solution will be unique. See [33] for more details on the representation of $V^h(\cdot, u)$.

A representation of $\gamma^h(u)$. By iterating (11.5), we get that for any $x \in G_h$, (use $u_i^h = u(\xi_i^h)$)

$$\gamma^h(u) = \lim_n \frac{E_x^u \sum_0^n k(\xi_i^h, u_i^h)\Delta t_i^h}{E_x^u \sum_0^n \Delta t_i^h}. \tag{11.6}$$

The limit exists by the ergodic theorem for Markov chains [8, § I.15]. We now rewrite (11.6) in a form that is suited to getting the limits for the cost for $\xi^h(\cdot)$. Let $\{\pi^h(x,u), x \in G_h\}$ denote the unique invariant measure for $\{\xi_n^h\}$, under the control $u(\cdot)$ and define the measure $\mu^h(\cdot, u)$ by

$$\begin{aligned} \mu^h(x,u) &= \frac{\Delta t^h(x)\pi^h(x,u)}{\sum_y \Delta t^h(y)\pi^h(y,u)}, \qquad x \in G_h, \\ &= 0, \qquad x \notin G_h. \end{aligned} \tag{11.7}$$

Then by the ergodic theorem for Markov chains [8, § I.15], we can rewrite (11.6) as

$$\gamma^h(u) = \sum_x k(x,u(x))\mu^h(x,u). \tag{11.8}$$

Also, by the ergodic theorem for Markov chains (the pathwise limits are w.p.1)

$$\begin{aligned} \gamma^h(u) &= \lim_n \left[\sum_0^n k(\xi_i^h, u_i^h)\Delta t_i^h / t_n^h \right] \\ &= \lim_n \frac{1}{t_n^h} \int_0^{t_n^h} k(\xi^h(s), u(\xi^h(s)))\, ds \\ &= \lim_t \frac{1}{t} \int_0^t k(\xi^h(s), u(\xi^h(s)))\, ds \\ &= \lim_t \frac{1}{t} E_x^u \int_0^t k(\xi^h(s), u(\xi^h(s)))\, ds. \end{aligned} \tag{11.9}$$

Throughout the above calculation, the times i at which $\xi_i^h \in \partial G_h^+$ do not appear, since the associated Δt_i^h is zero. For the purposes of the analysis to follow, it matters little whether or not we allow these states. For specificity in the development, for $\xi_n^h \in G_h$, we let ξ_{n+1}^h denote the next state which is in G_h; i.e., either the next true state—or the one obtained by the instantaneous reflection from the next true state.

Approximation to the invariant measure and average cost for $x(\cdot)$; fixed continuous feedback control $u(\cdot)$. The above calculations suggest that $\mu^h(\cdot, u)$ is an approximation to an invariant measure of $x(\cdot)$ under $u(\cdot)$, and that $\gamma^h(u)$ is an approximation to the average cost under $u(\cdot)$ and this will now be discussed.

An alternative representation of the interpolated chain. Suppose that $\xi^h(\cdot)$ is a continuous parameter Markov chain with control $u^h(\cdot)$ and transition probabilities $p^h(x, y \mid u^h)$ and *mean sojourn times* $\Delta t^h(x)$ at x. Let t_n^h denote the time of the nth state transition, and Δt_n^h the random variable, which is the nth sojourn time. Then (11.9) still holds [35], $\mu^h(\cdot, u)$ is the invariant measure for $\xi^h(\cdot)$, and

$$\gamma^h(u^h) = \int k(x, u^h(x))\mu^h(dx, u^h).$$

Let $\xi^h(\cdot)$ denote the *stationary* continuous parameter Markov process. Then

$$\gamma^h(u^h) = E^{u^h} \int_0^1 k(\xi^h(s), u^h(\xi^h(s)))\, ds.$$

Let $m^h(\cdot)$ denote the relaxed control representation of $u^h(\xi^h(\cdot))$. We can show that the weak limits $(x(\cdot), m(\cdot))$ of $\{\xi^h(\cdot), m^h(\cdot)\}$ are stationary, and that $x(\cdot)$ is a stationary process, which is of the type used in § 9 or 10 (as appropriate) driven by the control $m(\cdot)$. It follows that the limits of $\gamma^h(u^h)$ (for whatever feedback sequence $u^h(\cdot)$ is used) are average costs per unit time for some limit stationary process.

We can state the following result.

THEOREM 11.1. *Assume* (A2.1), (A2.2), (A11.1). *Then* $\overline{\lim}_h \gamma^h \leqq \gamma(u)$ *for any continuous feedback control* $u(\cdot)$ *for the system* $x(\cdot)$ *satisfying* (9.11) *or* (10.2) *for which there is a unique invariant measure.*

If there is a continuous feedback control $u(\cdot)$ *that is* δ*-optimal for* $x(\cdot)$, *then* $\underline{\lim}_h \gamma^h \geqq \bar{\gamma} - \delta$.

Remark. The theorem does not quite duplicate the result in Theorem 7.1. The basic reason concerns the difficulty in getting a large enough family of comparison controls. The invariant measure can be quite sensitive to the approximation—even if the behavior over a finite time interval is not. The class of comparison controls $u(\cdot)$ used in the theorem can be extended in many directions—provided only that there are approximations $\tilde{u}^h(\cdot)$ that can be applied to $\{\xi_n^h\}$ and yield limit costs $\gamma^h(\tilde{u}^h) \to \gamma(u)$. For the unreflected problem, it is shown in [36] that, under suitable conditions, any feedback control can be so approximated. This can also be done for the reflected problem, since the reflection direction is not controlled.

We could also use the following class. Let $u(\cdot)$ be discontinuous, with discontinuity set D_u. Let $N_\varepsilon(D_u)$ denote an ε-neighborhood of D_u. Suppose that for arbitrary control values used in $N_\varepsilon(D_u)$, the fraction (per unit time) of time $x(\cdot)$ (under $u(\cdot)$) spends in $N_\varepsilon(D_u)$ goes to zero as $\varepsilon \to 0$. Then $\overline{\lim}_h \gamma^h \leqq \gamma(u)$ also.

REFERENCES

[1] M. AKIAN, *Resolution numerique d'equations d'Hamilton-Jacobi-Bellman au moyen d'algorithmes multigrilles et d'iterations sur les politiques*, Eighth Conference on Analysis and Optimization of Systems (INRIA), Antibes, France, 1988.

[2] L. D. BERKOVITZ, *Optimal Control Theory*, Applied Mathematical Science, Vol. 12, Springer-Verlag, Berlin, 1974.

[3] D. P. BERTSEKAS AND D. A. CASTANON, *Adaptive aggregation methods for infinite horizon dynamic programming*, IEEE Trans. Automat. Control, 34 (1989), pp. 589-598.

[4] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[5] V. S. BORKAR AND M. K. GHOSH, *Ergodic control of multidimensional diffusions* I: *the existence results*, SIAM J. Control Optim., 26 (1988), pp. 112-126.

[6] P. CHANCELIER, C. GOMEZ, J.-P. QUADRAT, A. SULEM, G. L. BLANKENSHIP, A. LA VIGNA, D. C. MACENARY, AND I. YAN, *An expert system for control and signal processing with automatic* FORTRAN *program generation*, Mathematical Systems Symposium, Stockholm, Royal Institute of Technology, Stockholm, Sweden, 1986.

[7] P. CHANCELIER, C. GOMEZ, J.-P. QUADRAT, AND A. SULEM, *Automatic study in stochastic control*, IMA Volumes in Mathematics and Its Applications, Vol. 10, W. Fleming and P. L. Lions, eds., Springer-Verlag, Berlin, 1987.

[8] K. L. CHUNG, *Markov Chains with Stationary Transition Probabilities*, Springer-Verlag, Berlin, 1960.

[9] C. A. CLAROTTI, G. F. PUGNO AND W. J. RUNGGALDIER, *Statistical reliability practice from sampling theory to stochastic filtering*, Reliability Engineering and System Safety, 26 (1989), pp. 1-20.

[10] M. COX AND I. KARATZAS, *Stationary and discounted control for diffusions*, in Proc. 3rd Bad Honnef Meeting on Stochastic Systems, Lecture Notes in Control and Information Sciences, Vol. 78, Springer-Verlag, Berlin, 1985.

[11] G. B. DIMASI AND W. J. RUNGGALDIER, *An approximation method for nonlinear filtering*, S. K. Mitter and A. Moro, eds., Lecture Notes in Mathematics 972, Springer-Verlag, Berlin, 1982, pp. 249-259.

[12] ———, *Approximations and bounds for discrete time non-linear filtering*, Lecture Notes in Control and Information Sciences, Vol. 44, Springer-Verlag, Berlin, 1982.

[13] E. B. DYNKIN, *Markov Processes*, Springer-Verlag, Berlin, 1965.

[14] N. EL KAROUI AND S. MELEARD, *Martingale measures and stochastic calculus*, preprint, Lab. de Probabilities, University of Paris VI, Paris, France, 1987.

[15] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.

[16] M. FALCONE, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), p. 1-13.

[17] W. H. FLEMING, *Generalized solutions in optimal stochastic control*, in Proc. U.R.I. Conference on Control, University of Rhode Island, Providence, RI, 1982.

[18] W. H. FLEMING AND M. NISIO, *On stochastic relaxed controls for partially observed diffusions*, Nagoya Math. J., 93 (1984), pp. 71-108.

[19] A. GERMIANI AND M. PICCIONI, *Some discretizations of stochastic partial differential equations on* R^d *by a finite element method*, Stochastics, 23 (1988), pp. 131-139.

[20] I. I. GIHMAN AND A. V. SKOROHOD, *Introduction to the Theory of Random Processes*, W. B. Saunders, Philadelphia, 1965.

[21] R. GONZALEZ AND E. ROFMAN, *On deterministic control problems: an approximation procedure for the optimal cost, parts* I, II, SIAM J. Control Optim., 23 (1985), pp. 242-285.

[22] J. M. HARRISON, *Brownian Motion and Stochastic Flow Systems*, John Wiley, New York, 1985.

[23] J. M. HARRISON AND M. I. REIMAN, *Reflected Brownian motion on an orthant*, Ann. Probab., 9 (1981), pp. 302-308.

[24] D. L. IGLEHART AND W. WHITT, *Multiple channel queues in heavy traffic*, Adv. in Appl. Probab., 2 (1970), pp. 150-177.

[25] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion processes.* North-Holland, Amsterdam, 1981.

[26] I. KARATZAS, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15 (1983) 5 pp. 225-254.

[27] I. KARATZAS AND S. E. SHREVE, *Equivalent models for finite fuel stochastic control*, Stochastics, 18 (1986), pp. 245-276.

[28] T. G. KURTZ, *Approximation of Population Processes*, CBMS-NSF Regional Conference Series in Applied Mathematics 36, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1981.

[29] H. J. KUSHNER, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[30] ———, *A robust computable approximation to the optimal nonlinear filter*, Stochastics, 3 (1979), pp. 75-83.

[31] H. J. KUSHNER AND K. M. RAMACHANDRAN, *Optimal and approximately optimal control policies for queues in heavy traffic*, Lefschetz Center for Dynamical Systems Report Brown University, Providence, RI, 1987; SIAM J. Control Optim., (1989), pp. 1293-1318.

[32] H. J. KUSHNER AND A. J. KLEINMAN, *Accelerated procedures for the solution of discrete Markov control problems*, IEEE Trans. Automat. Control, 16 (1971), pp. 147-152.

[33] H. J. KUSHNER, *Introduction to Stochastic Control Theory*, Holt, Reinhart and Winston, New York, 1972.

[34] ———, *Probabilistic methods for finite difference approximations to degenerate elliptic and parabolic equations with Neumann and Dirichlet boundary conditions*, J. Math. Appl., 53 (1976), pp. 644-668.

[35] H. J. KUSHNER AND G. B. DIMASI, *Approximations for functionals and optimal control problems on jump-diffusion processes*, J. Math. Anal. Appl., 63 (1978), pp. 772-800.

[36] H. J. KUSHNER, *Optimality conditions for the average cost per unit time problem with a diffusion model*, SIAM J. Control Optim., 16 (1978), pp. 330–346.

[37] H. J. KUSHNER AND W. RINGGALDIER, *Nearly optimal state feedback controls for stochastic systems with wideband noise disturbances*, SIAM J. Control Optim., 25 (1987), pp. 289–315.

[38] H. J. KUSHNER AND K. M. RAMACHANDRAN, *Nearly optimal singular controls for wideband noise driven systems*, SIAM J. Control Optim., 26 (1988), pp. 569–591.

[39] A. J. LEMOINE, *Networks of queues: a survey of weak convergence results*, Management Sci., 24 (1978), pp. 1175–1193.

[40] P. L. LIONS AND A. S. SZNITMAN, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 53 (1984), pp. 644–668.

[41] J. L. MENALDI, *Some estimates for finite difference approximations*, SIAM J. Control Optim., 27 (1989), pp. 579–607.

[42] M. L. PUTERMAN, *Markov decision processes*, Ch. 8, Handbook of Operations Research and Management Science, Vol. 2, D. P. Heyman, M. J. Sobel, eds., Elsevier (North-Holland), 1990, pp. 331–433.

[43] J.-P. QUADRAT, *Sur l'identification et le controle de systems dynamiques stochastiques*, Thesis, University of Paris, Paris, France, 1981.

[44] M. I. REIMAN, *Open queueing networks in heavy traffic*, Math. Oper. Res., 9 (1984), pp. 441–458.

[45] P. J. SCHWEITZER, M. PUTERMAN, AND K. W. KINDLE, *Iterative aggregation-disaggregation procedures for solving discounted semi-Markovian reward processes*, Oper. Res., 33 (1985), pp. 589–606.

[46] P. J. SCHWEITZER, *Solving Markovian decision processes by successive elimination of variables*, J. Math. Anal. Appl., 103 (1988), pp. 403–419.

[47] D. W. STROOCK AND S. R. S. VARADHAN, *Diffusion processes with boundary conditions*, Comm. Pure Appl. Math., 24 (1971), pp. 147–225.

[48] ———, *On degenerate elliptic and parabolic operators of second order and their associated diffusions*, Comm. Pure Appl. Math., 25 (1972), pp. 651–713.

[49] ———, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, 1979.

[50] J. WARGA, *Relaxed variational problems*, J. Math. Appl., 4 (1962), pp. 111–128.

[51] J. B. WALSH, *An introduction to stochastic partial differential equations*, Ecole d'ete de probabilities de Saint Flour XIV, Springer-Verlag Berlin, 1984.

[52] D. J. WHITE, *Markov decision models for the evaluation of a large class of continuous sampling plans*, Ann. Math. Statist., 36 (1965), pp. 1408–1420.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 1049-1060, September 1990

002

IDENTIFICATION OF DISCONTINUOUS PARAMETERS IN FLOW EQUATIONS*

S. GUTMAN†

Abstract. A problem of identifying possibly discontinuous diffusion coefficients in parabolic equations is considered. General theorems on existence and convergence of Galerkin approximations are proved in L^1 setting. Classes of functions of bounded variation are discussed and the variation estimates are obtained. A double-discretization method with the variations constraints is used in two- and three-dimensional problems and the numerical experiments are presented.

Key words. parameters identification, inverse problem, flow equations, Galerkin approximations

AMS(MOS) subject classification. 35R30

1. Introduction.

Let Ω be a bounded open domain in $\mathbb{R}^3$. Then the system

$$\begin{aligned} &u_t - \nabla(a(x)\nabla u) = f(x,t), && (x,t)\in\Omega\times(0,T),\\ &u(x,t)=0, && (x,t)\in\partial\Omega\times(0,T),\\ &u(x,0)=u_0(x), && x\in\Omega, \end{aligned} \tag{1}$$

describes a flow of a fluid through the medium with permeability $a(x)$, $x\in\Omega$. We will assume that $a\in L^\infty(\Omega)$ and, moreover, $a\in A_{\text{ad}}$,

$$A_{\text{ad}}=\{a(x)\in L^\infty(\Omega)\colon 0<\nu\leqq a(x)\leqq\mu \text{ a.e. in } \Omega\}. \tag{2}$$

Condition (2) reflects the fact that in a physically relevant situation the permeability $a(x)$ is assumed to be taken between the prescribed bounds ν and μ. If $u_0\in L^2(\Omega)$ and $f\in L^2(0,T;L^2(\Omega))$, then it is well known (see, e.g., [1, Chap. 3], [2, Chap. 3]), that the system (1), (2) has a unique (weak) solution $u(x,t)$ that we will also denote as $u(a)$ to emphasize its dependence on the coefficient $a(x)\in A_{\text{ad}}$. This solution is an element of $C([0,T];L^2(\Omega))$.

The parameter estimation problem for (1) and (2) seeks to determine the coefficient $a(x)$ in such a way that the solution $u(a)$ "matches" the observed flow $z(x,t)$ of (1) in a prescribed sense (see [3]-[10] for general information).

To be precise we say that the coefficient $\bar{a}\in K_{\text{ad}}\subset A_{\text{ad}}$ solves the parameter estimation problem for the admissible set K_{ad} if

$$J(\bar{a})=\inf\{J(a)\colon a\in K_{\text{ad}}\}, \tag{3}$$

where $J(a)=\|u(a)(T)-z(T)\|_{L^2(\Omega)}$.

Thus we must determine if there is a solution for (1)-(3) and how this solution is related to its approximations, which are obtained in a process of numerical computations.

It can be shown (see Theorem 3.2 and [11]) that the mapping $a\to u(a)$ is continuous considered from $A_{\text{ad}}\subset L^2(\Omega)$ into $C([0,T];L^2(\Omega))$. Thus (1)-(3) has a solution if the admissible set of parameters $K_{\text{ad}}\subset A_{\text{ad}}$ is taken to be compact in $L^2(\Omega)$. Indeed, this is the argument in many papers (see, e.g., [6], [8]) where it is reasonable to assume sufficient smoothness of the involved coefficients $a(x)$. The set $K_{\text{ad}}=\{a\in H^1(\Omega)\colon \|a\|_{H^1}\leqq\text{const.}\}$ is compact in $L^2(\Omega)$ and the functional $J(a)=\|u(a)-z\|$ attains a minimum on it.

* Received by the editors May 11, 1987; accepted for publication (in revised form) October 27, 1988.
† Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019. This research was partially supported by Department of Energy grant DE-FG01-87FE61146.

Of course, this solution $\bar{a}$ is in $H^1(\Omega)$. However the subject of this paper is the identification of parameters in (1)–(3) that refer to a different physical situation. Suppose, for example, that $\Omega = (0, 1) \times (0, 1)$, $\Omega_1 = (0, \frac{1}{2}) \times (0, 1)$, and $\Omega_2 = (\frac{1}{2}, 1) \times (0, 1)$. $a(x) = \mu$ if $x \in \Omega_1$, and $a(x) = \nu$ if $x \in \Omega_2$. Thus the medium consists of two regions with different permeabilities and we would like to develop a method for identification of discontinuous coefficients $a(x)$ in Ω. Therefore, the admissible set K_{ad} must include the coefficients of the type described above. However this coefficient $a(x)$ does not belong to $H^1(\Omega)$ and therefore a different set K_{ad} should be considered. Before we begin a systematic treatment of the subject, let us mention some alternate approaches.

In [9] and [10] the assumptions imply the continuity and differentiability of the coefficient $a(x)$. In [12] the methods for the detection of discontinuities are not fully developed for two- and three-dimensional problems.

In [13], Gutman and White consider an approach based on G-convergence of parabolic operators. In this method the admissible set of parameters K_{ad} is taken to be $A_{\text{ad}} = \{a \in L^\infty(\Omega): 0 < \nu \leqq a(x) \leqq \mu, \text{ a.e. on } \Omega\}$. Introducing $d(a_1, a_2)$ on A_{ad} by $d(a_1, a_2) = \|u(a_1) - u(a_2)\|_{L^2(Q)}$, $Q = (0, T) \times \Omega$, the set A_{ad} becomes a precompact set (of classes) of coefficients. The theory of G-convergence [11], [14] shows that its completion can be achieved by embedding the set A_{ad} into the set of all second-order elliptic operators. Thus, in this approach, we should question how appropriate is the studied mathematical model for the interpretation of the physical reality.

2. Functions of bounded variation. Let $\Omega \subset \mathbb{R}^n$ be an open bounded set in $\mathbb{R}^n$ with a Lipschitz continuous boundary $\partial\Omega$. By $|x|$ we denote $|x| = (\sum_{i=1}^n x_i^2)^{1/2}$ for $x = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$. Following [15] define the variation $\int_\Omega |Df|$ of a function $f \in L^1(\Omega)$ as

$$\int_\Omega |Df| = \sup \left\{ \int_\Omega f \operatorname{div} g \, dx : g = (g_1, \cdots, g_n) \in C_0^1(\Omega; \mathbb{R}^n) \text{ and } |g(x)| \leqq 1 \text{ for } x \in \Omega \right\}, \tag{4}$$

where $\operatorname{div} g = \sum_{i=1}^n (\partial g_i / \partial x_i)$.

If the variation of f is finite, that is, $\int_\Omega |Df| < \infty$, we say that f has a bounded variation. The space of all functions $f \in L^1(\Omega)$ with bounded variation is denoted by $BV(\Omega)$.

Example 2.1. If $f \in C^1(\Omega)$, then $\int_\Omega f \operatorname{div} g \, dx = -\int_\Omega \sum_{i=1}^m (\partial f / \partial x_i) g_i \, dx$ for every $g \in C_0^1(\Omega, \mathbb{R}^n)$ and $\int_\Omega |Df| = \int_\Omega |\operatorname{grad} f| \, dx$ where $\operatorname{grad} f = (\partial f / \partial x_1, \partial f / \partial x_2, \cdots, \partial f / \partial x_n)$ and Ω is assumed to satisfy the conditions of the divergence (Gauss–Green) theorem.

Example 2.2. Let B be a ball in Ω, and let $\chi(B)$ be its characteristic function. Then $\int_\Omega \chi(B) \operatorname{div} g \, dx = \int_B \operatorname{div} g \, dx = \int_{\partial B} g \cdot \nu \, ds$ for any $g \in C_0^1(\Omega, \mathbb{R}^n)$, where ν is the outward unit normal to ∂B. Thus $\int_\Omega |D\chi| \leqq |\partial B|$—the surface area of B. (In fact, the variation of χ is equal to $|\partial B|$ in this case.)

The most important properties of the space $BV(\Omega)$ are the following.

PROPOSITION 2.3 [15, Thm. 1.9] (semicontinuity). *If* $\{f_j\}_{j=1}^\infty \subset BV(\Omega)$ *and* $f_j \to f$ *as* $j \to \infty$ *in* $L^1(\Omega)$, *then* $\int_\Omega |Df| \leqq \lim_{j \to \infty} \inf \int_\Omega |Df_j|$.

PROPOSITION 2.4 [15, Thm. 1.12]. *Under the norm* $\|f\|_{BV} = \|f\|_{L^1} + \int_\Omega |Df|$, $BV(\Omega)$ *is a Banach space.*

PROPOSITION 2.5 [15, Thm. 1.17] (approximation). *Let* $f \in BV(\Omega)$. *Then there exists a sequence* $\{f_j\}_{j=1}^\infty \in C^\infty(\Omega)$ *such that* $\lim_{j \to \infty} \int_\Omega |f - f_j| \, dx = 0$ *and* $\lim_{j \to \infty} \int_\Omega |Df_j| = \int_\Omega |Df|$.

Finally, we have the following proposition.

PROPOSITION 2.6 [15, Thm. 1.19] (compactness). *Sets of functions uniformly bounded in BV norm are relatively compact in* $L^1(\Omega)$.

COROLLARY 2.7. *The set* $K_c = \{f \in A_{\text{ad}} : \int_\Omega |Df| \leqq C\}$ *is compact in* $L^1(\Omega)$ *for any* $C > 0$.

Proof. If $\int_\Omega |Df| \leqq C$, then $\|f\|_{BV} \leqq C + \mu|\Omega|$. Thus K_c is precompact in $L^1(\Omega)$ by the previous proposition. If $f_j \to f$ in $L^1(\Omega)$ as $j \to \infty$ and $\int_\Omega |Df_j| \leqq C$, then $\int_\Omega |Df| \leqq C$ by Proposition 2.3 and K_c is closed in $L^1(\Omega)$.

Thus to use the above theory of functions of bounded variation for numerical computations, we must estimate variation of functions with rather general discontinuities. To this end let us introduce the following hypothesis:

(H) Let $\Omega, \Omega_i, 1 \leqq i \leqq p$ be open, bounded sets of $\mathbb{R}^n$; $\Omega_i \subset \Omega$, $1 \leqq i \leqq n$, $\Omega_i \cap \Omega_j = \varnothing$ for $i \neq j$ and $\cup_{i=1}^p \bar{\Omega}_i \supset \Omega$. Let the boundaries $\partial\Omega, \partial\Omega_i$, $1 \leqq i \leqq p$ be Lipschitz continuous and each Ω_i satisfy the conditions of the divergence (Gauss–Green) theorem.

To state our next theorem let Ω, Ω_i, $1 \leqq i \leqq p$ satisfy (H), and let f be a function from $L^1(\Omega)$ such that its restrictions $f|_{\Omega_i}$ on each of Ω_i, $1 \leqq i \leqq p$ are continuous and can be continuously extended to $\bar{\Omega}_i$. Let Γ be a common boundary of Ω_i and Ω_j. By $|f(\Gamma)|$ we will understand the absolute value of the difference $\tilde{f}_i|_\Gamma - \tilde{f}_j|_\Gamma$ where $\tilde{f}_i$ is the extension of f to $\bar{\Omega}_i$ and $\tilde{f}_j$ is the extension of f to $\bar{\Omega}_j$.

THEOREM 2.8. *Let* Ω, Ω_i, $1 \leqq i \leqq p$ *satisfy hypothesis* (H), *let function* $f \in L^1(\Omega)$ *be continuously differentiable on* Ω_i, $1 \leqq i \leqq p$ *and can be continuously extended in every* $\bar{\Omega}_i$, $1 \leqq i \leqq p$. *Let* Γ_j, $1 \leqq j \leqq m$ *be common boundaries of the domains* Ω_i, $1 \leqq i \leqq p$. *Then*

$$\int_\Omega |Df| \leqq \sum_{i=1}^p \int_{\Omega_i} |\operatorname{grad} f|\, dx + \sum_{j=1}^m \int_{\Gamma_j} |f(\Gamma_j)|\, ds. \tag{5}$$

Proof. Define $f_i(x) = f(x)$ for $x \in \Omega_i$, $1 \leqq i \leqq p$ and $f_i(x) = 0$ for $x \notin \bar{\Omega}_i$. We will write $\tilde{f}_i(x)$ for the continuous extension of f_i to $\bar{\Omega}_i$ (from Ω_i). Thus $f = \sum_{i=1}^p f_i$. For every f_i we have

$$\int_\Omega f_i \operatorname{div} g\, dx = -\int_{\Omega_i} (\operatorname{grad} f_i) \cdot g\, dx + \int_{\partial\Omega_i} \tilde{f}_i g \cdot \nu_i\, ds,$$

where $g \in C_0^1(\Omega; \mathbb{R}^n)$ and ν_i is the unit outward normal to $\partial\Omega_i$. For the sum $f = \sum_{i=1}^p f_i$ we obtain

$$\begin{aligned}\int_\Omega f \operatorname{div} g\, dx &= -\sum_{i=1}^p \int_{\Omega_i} \operatorname{grad} f \cdot g\, dx + \sum_{i=1}^p \int_{\partial\Omega_i} \tilde{f}_i g \cdot \nu_i\, ds \\ &= -\sum_{i=1}^p \int_{\Omega_i} \operatorname{grad} f \cdot g\, dx + \sum_{j=1}^m \int_{\Gamma_j} (\tilde{f}_l \nu_l + \tilde{f}_k \nu_k) \cdot g\, ds,\end{aligned}$$

where Γ_j is the common boundary of the domains Ω_l and Ω_k. Note also that $\nu_k = -\nu_l$. Thus $\left|\int_\Omega f \operatorname{div} g\, dx\right| \leqq \sum_{i=1}^p \int_{\Omega_i} |\operatorname{grad} f|\,|g|\, dx + \sum_{j=1}^m \int_{\Gamma_j} |f(\Gamma_j)|\,|g|\, ds$ and $\int_\Omega |Df| \leqq \sum_{i=1}^p \int_{\Omega_i} |\operatorname{grad} f|\, dx + \sum_{j=1}^m \int_{\Gamma_j} |f(\Gamma_j)|\, ds$.

3. Galerkin approximations and the continuity results. Let $\{\phi_i\}_{i=1}^N$ be a linearly independent set in $V = H_0^1(\Omega)$. Given a coefficient $a(x) \in A_{\text{ad}}$, we define the Galerkin approximation $u^N(x, t)$ (with respect to the basis $\{\phi_i\}_{i=1}^N$) as $u^N(x, t) = \sum_{i=1}^N c_i(t)\phi_i(x)$ where $\tilde{c} = \{c_i(t)\}_{i=1}^N$ is determined as the solution of the matrix equation

$$G\tilde{c}' + \hat{G}(a)\tilde{c} = \tilde{f}, \qquad G\tilde{c}(0) = \tilde{c}_0, \tag{6}$$

where G is the stiffness matrix

$$\{G\}_{ij} = \int_\Omega \phi_i \phi_j \, dx, \qquad \{\hat{G}(a)\}_{ij} = \int_\Omega a(x) \nabla \phi_i \nabla \phi_j \, dx,$$

$$\{\tilde{f}\}_i = \int_\Omega f(x, t) \phi_i(x) \, dx, \qquad \{\tilde{c}_0\}_i = \int_\Omega u_0(x) \phi_i(x) \, dx.$$

Equation (6) is obtained (see, e.g., [1]) from the conditions

$$(u_t^N, \phi_i) + (a \nabla u^N, \nabla \phi_i) = (f, \phi_i), \qquad 1 \leqq i \leqq N,$$

$$u^N(0) = \sum_{i=1}^N \xi_i^N \phi_i, \quad \sum_{i=1}^N \xi_i^N \phi_i \to u_0 \quad \text{in } L^2(\Omega) \quad \text{as } N \to \infty, \tag{7}$$

where $(\cdot, \cdot)$ is the dot product in $H = L^2(\Omega)$. Let V' be the dual to $V = H_0^1(\Omega)$; then $V' = H^{-1}(\Omega)$. The standard methods [1], [2] imply that if $\operatorname{Span}\{\phi_i\}_1^\infty$ is dense in V, then $u^N(a) \to u(a)$, strongly in $L^2(0, T; V)$ as $N \to \infty$ and the functions $u(a)$, $u^N(a)$ can be considered as elements of $C([0, T]; H)$. Moreover (see the Appendix), $u^N(a) \to u(a)$ in $C([0, T]; H)$ as $N \to \infty$.

LEMMA 3.1. *Let Z be a compact set in $L^1(\Omega)$ and $\hat{M} > 0$. Then*

$$\eta_Z(\delta) = \sup \left\{ \left| \int_\Omega gh \, dx \right| : \|g\|_1 \leqq \delta, \|g\|_\infty \leqq \hat{M}, h \in Z \right\} \to 0$$

as $\delta \to 0$.

Proof. Given $\varepsilon > 0$, there exists a finite set $K_\varepsilon \subset L^1(\Omega)$ such that $\|f\|_\infty \leqq C_\varepsilon$ for all $f \in K_\varepsilon$ and $\min\{\|h - f\|_1 : f \in K_\varepsilon\} \leqq \varepsilon$ for any $h \in Z$. Let $0 < \delta < \varepsilon / C_\varepsilon$. Then $|\int_\Omega gh \, dx| \leqq \int_\Omega |gf| \, dx + \int_\Omega |g| \, |f - h| \, dx \leqq C_\varepsilon \|g\|_1 + \hat{M} \|f - h\|_1$ for every $g \in L^1(\Omega)$ with $\|g\|_\infty \leqq \hat{M}$ and $f \in K_\varepsilon$. If $\|g\|_1 \leqq \delta$ and $f \in K_\varepsilon$ is chosen appropriately, we get $|\int_\Omega gh \, dx| \leqq \varepsilon + \hat{M}\varepsilon$ and the lemma is proved.

THEOREM 3.2. (a) *The mappings $u^N, u : L^1(\Omega) \to C([0, T]; H)$ are continuous on A_{ad}.*

(b) *The convergence $u^N(a) \to u(a)$ in $C([0, T]; H)$ is uniform on any compact set $K \subset L^1(\Omega)$.*

Proof. Let $|\cdot|, \|\cdot\|_V, \|\cdot\|_{V'}$ be the norms in $H = L^2(\Omega)$, $V = H_0^1(\Omega)$, and $V' = H^{-1}(\Omega)$ correspondingly. By [1, § 3.1.4]

$$\int_0^T \|u^N(x, t)\|_V^2 \, dt \leqq c \left(|u_0|^2 + \int_0^T \|f(x, t)\|_{V'}^2 \, dt \right), \tag{8}$$

$N = 1, 2, \cdots$. The same inequality also holds for the (weak) solution $u(x, t)$ of (1), where the coefficient $a(x) \in A_{\text{ad}}$.

If $\{a_n\}_1^\infty$, $a \in A_{\text{ad}}$ and $u^N(a_n) = \sum_{i=1}^N c_{ni}(t) \phi_i(x)$, then, from (7) $(u_t^N(a_n) - u_t^N(a), \phi_i) + (a_n[\nabla u^N(a_n) - \nabla u^N(a)], \nabla \phi_i) = ((a - a_n) \nabla u^N(a), \nabla \phi_i)$, $1 \leqq i \leqq N$. Multiplying this equality by $c_{ni}(t) - c_i(t)$ and taking the sum from $i = 1$ to N, we get

$$\begin{aligned} &([u^N(a_n) - u^N(a)]_t, u^N(a_n) - u^N(a)) \\ &\quad + (a_n[\nabla(u^N(a_n) - u^N(a))], \nabla(u^N(a_n) - u^N(a))) \\ &\quad = ((a - a_n) \nabla u^N(a), \nabla(u^N(a_n) - u^N(a))), \end{aligned}$$

or writing the first term as $\frac{1}{2} d/dt |u^N(a_n) - u^N(a)|^2$ and integrating both sides from

zero to T, we obtain

$$|(u^N(a_n)-u^N(a))(T)|^2 = -2\int_0^T\int_\Omega a_n(x)|\nabla(u^N(a_n)-u^N(a))|^2\,dx\,dt$$
$$+2\int_0^T\int_\Omega (a-a_n)\nabla u^N(a)\nabla(u^N(a_n)-u^N(a))\,dx\,dt,$$

where we have used $u^N(a_n)(0)=u^N(a)(0)$. Thus

$$|(u^N(a_n)-u^N(a))(T)|^2 \leqq 2\int_0^T\int_\Omega (a-a_n)\nabla u^N(a)\nabla(u^N(a_n)-u^N(a))\,dx\,dt.$$

Finally, this inequality and (8) gives

$$|(u^N(a_n)-u^N(a))(T)|^2 \leqq C\left[\int_0^T\int_\Omega |a-a_n|^2|\nabla u^N(a)|^2\,dx\,dt\right]^{1/2} \tag{9}$$

Similarly,

$$|(u(a_n)-u(a))(T)|^2 \leqq C\left[\int_0^T\int_\Omega |a-a_n|^2|\nabla u(a)|^2\,dx\,dt\right]^{1/2} \tag{10}$$

Let $a_n \to a$ in $L^1(\Omega)$ as $n\to\infty$. Since $\|a\|_1 \leqq$ const. $\|a\|_2$ and $\|a\|_2 \leqq \|a\|_1\|a\|_\infty$, the topologies of $L^1(\Omega)$ and $L^2(\Omega)$ coincide on A_{ad}. Thus $\int_\Omega |a_n-a|^2\,dx\to 0$ as $n\to\infty$.

Fix $t\in[0,T]$. Define $g_n\in L^1(\Omega)$ and $h\in L^1(\Omega)$ by $g_n(x)=|a_n(x)-a(x)|^2$ and $h(x)=|\nabla u(x)|^2$. By Lemma 3.1, $\int_\Omega g_n(x)h(x)\,dx\to 0$ as $n\to\infty$. By the Lebesgue dominated convergence theorem applied to $G_n(t)=\int_\Omega |a-a_n|^2|\nabla u(a)|^2\,dx$, the right-hand side in (10) goes to zero as $n\to\infty$ and $u(a_n)\to u(a)$ in $C([0,T];H)$ as $n\to\infty$. Similarly, $u^N(a_n)\to u^N(a)$ in $C([0,T];H)$.

To prove (b) of the theorem, we note that $u^N(a)\to u(a)$ strongly in $L^2(0,T;V)$ as $N\to\infty$ (see [1]). As shown in the Appendix, $u^N(a)\to u(a)$ in $C([0,T];H)$ for each $a\in A_{\text{ad}}$, therefore it is sufficient to show that the mappings $u^N(a):A_{\text{ad}}\to C([0,T];H)$, $n=1,2,\cdots$ are equicontinuous on A_{ad}. Let $a\in A_{\text{ad}}$ and $\varepsilon>0$ be given. Since $u^N(a)\to u(a)$, $N\to\infty$ in $L^2(0,T;V)$, the set $\{u^N(a)\}_{N=1}^\infty$ is precompact in $L^2(0,T;V)$. Therefore, there exists a finite set $Q_\varepsilon\subset L^2(0,T;V)$ such that any member g of Q_ε is a continuous function from $[0,T]$ to V and $\min\{\|u^N(a)-g\|_{L^2(0,T;V)}: g\in Q_\varepsilon\}\leqq\varepsilon$ for a given N. It follows from the continuity of the functions $g\in Q_\varepsilon$ that the range $Z=\cup\{|\nabla g(t)|^2: g\in Q_\varepsilon, t\in[0,T]\}$ is compact in $L^1(\Omega)$. Let $\delta>0$, $b\in A_{\text{ad}}$ and $\|a-b\|_1\leqq\delta$.

From (9) we obtain

$$\begin{aligned}|(u^N(a)-u^N(b))(t)|^2 &\leqq C\|(a-b)|\nabla u^N(a)|\|_{L^2(0,T;H)}\\ &\leqq C\|(a-b)|\nabla u^N(a)-\nabla g|\|_{L^2(0,T;H)}+C\|(a-b)|\nabla g|\|_{L^2(0,T;H)}\\ &\leqq 2C\mu\varepsilon+C\left[\int_0^T\int_\Omega |a-b|^2|\nabla g|^2\,dx\,dt\right]^{1/2}\\ &\leqq 2C\mu\varepsilon+C\sqrt{T}\sqrt{\eta_Z(\delta)},\end{aligned}$$

where $\eta_Z(\delta)$ is as defined in Lemma 3.1 and C is independent of N.

This lemma shows that $\eta_Z(\delta)\to 0$ as $\delta\to 0$, and we conclude that $\|u^N(a)-u^N(b)\|^2_{C([0,T];H)}\leqq 2CM\varepsilon+\varepsilon$ for sufficiently small δ. Thus the functions $a\to u^N(a)$, $N=1,2,\cdots$ are equicontinuous, and the pointwise on A_{ad} convergence $u^N(a)\to u(a)$, $N\to\infty$ in $C([0,T];H)$ is, in fact, uniform on every compact set $K\subset L^1(\Omega)$ (that is, $u^N\to u$ in $C(K;C([0,T];H))$) and the theorem is proved.

4. Existence of the optimal coefficients. Theorem 3.2 from the previous section states that the mapping $a \to u(a)$ is continuous from $L^1(\Omega)$ to $C([0, T]; L^2(\Omega))$. The cost functional

$$J(a) = \|(u(a) - z)(T)\|_{L^2(\Omega)}$$

is therefore continuous on $A_{\text{ad}} \subset L^1(\Omega)$. Hence it attains its minimum on any compact set $K_{\text{ad}} \subset A_{\text{ad}}$. Combining this with Proposition 2.6 and Corollary 2.7, we obtain Theorem 4.1.

THEOREM 4.1. *The parameter estimation problem* (1)-(3) *has a solution on any admissible set* K_{ad} *of the form* $K_{\text{ad}} = \{a \in A_{\text{ad}}: \int_\Omega |Da| \leqq \text{const.}\}$.

That is, this theorem states that there exists $\bar{a} \in K_{\text{ad}}$ such that $J(\bar{a}) = \inf\{J(a): a \in K_{\text{ad}}\}$.

To find the coefficient $\bar{a}$ numerically we use a double-discretization algorithm, where both state and parameter variables u and a are approximated in finite-dimensional spaces H^N and A^M, respectively. Relations between the solutions of these approximation problems and the solutions of (1)-(3) are summarized in the following theorem.

THEOREM 4.2. *Let* $A_{\text{ad}}^M \subset A_{\text{ad}}$, $M \to \infty$ *be an increasing sequence of closed in* $L^1(\Omega)$ *subsets of* A_{ad}. *Let* $K \subset A_{\text{ad}}$ *be a compact set in* $L^1(\Omega)$, $K \cap A_{\text{ad}}^M \neq \varnothing$, *and let* $\bigcup_M (K \cap A_{\text{ad}}^M)$ *be dense in* K. *Let natural sequences* $M = M(q) \to \infty$ *and* $N = N(q) \to \infty$ *as* $q \to \infty$. *Also let*

$$J^q(a) = \|(u^N(a) - z)(T)\|_{L^2(\Omega)},$$

where the Galerkin approximations u^N *are defined as in* § 3. *Then*

(a) *There exists* $\bar{a}_q \in K \cap A_{\text{ad}}^M$ *such that* $J^q(\bar{a}_q) = \inf\{J^q(a): a \in A_{\text{ad}}^M \cap K\}$.

(b) *Any cluster point* $\bar{a}$ *of the sequence* $\{\bar{a}_q\}_{q=1}^\infty$ *satisfies* $J(\bar{a}) = \inf\{J(a): a \in K\}$.

(c) $\lim_q J^q(\bar{a}_q) = \inf\{J(a): a \in K\} = \lim_q J(\bar{a}_q)$.

(d) *If there exists* $\tilde{a} \in K$ *such that* $z = u(\tilde{a})$, *then* $\lim_q J^q(\bar{a}_q) = \lim_q J(\bar{a}_q) = 0$.

Proof. Theorem 3.2 shows that the functionals J^q, J are continuous on K and the convergence $J^q(a) \to J(a)$, $q \to \infty$ is uniform on K. Since $K \cap A_{\text{ad}}^M$ is compact, the functional J^q attains its minimum on it and (a) is done. To show (b) let $\{b_q\}_{q=1}^\infty \subset K$. If $\{b_{q_i}\}_{i=1}^\infty$ is a convergent subsequence and $b_{q_i} \to \bar{b}$ as $i \to \infty$, then $\lim_{i\to\infty} J^{q_i}(b_{q_i}) = J(\bar{b})$, since $|J^{q_i}(b_{q_i}) - J(\bar{b})| \leqq |J^{q_i}(b_{q_i}) - J(b_{q_i})| + |J(b_{q_i}) - J(\bar{b})|$. Therefore, if $\{\bar{a}_{q_i}\}_{i=1}^\infty \subset \{a_q\}_{q=1}^\infty$ and $\lim_{i\to\infty} \bar{a}_{q_i} = \bar{a}$, then $\lim_{i\to\infty} J^{q_i}(\bar{a}_{q_i}) = J(\bar{a})$. Let $\hat{a} \in K$ be such that $J(\hat{a}) = \inf\{J(a): a \in K\}$. Since $\bigcup_M (K \cap A_{\text{ad}}^M)$ is dense in K there exists a sequence $\{b_q\}_1^\infty$ such that $b_q \in K \cap A_{\text{ad}}^M$, $(M = M(q))$ and $b_q \to \hat{a}$ as $q \to \infty$. Then $J^{q_i}(\bar{a}_{q_i}) \leqq J^{q_i}(b_{q_i})$. Since $\lim_i J^{q_i}(b_{q_i}) = J(\hat{a})$ and $\lim_i J^{q_i}(\bar{a}_{q_i}) = J(\bar{a})$, we get $J(\bar{a}) \leqq J(\hat{a})$. Thus $J(\bar{a}) = J(\hat{a})$ and part (b) is proved.

Let $\gamma_q = J^q(\bar{a}_q)$, $q = 1, 2, \cdots$. The functions $\{J^q(a)\}_{q=1}^\infty$ are equicontinuous on K, hence the real sequence $\{\gamma_q\}_{q=1}^\infty$ is bounded and we have just shown that $\gamma = \inf\{J(a): a \in K\}$ is its cluster point. To see that γ is the unique cluster point of $\{\gamma_q\}_{q=1}^\infty$ let $\gamma^* = \lim_{i\to\infty} \gamma_{q_i}$. Without loss of generality (passing to a subsequence) we can assume that the correspondent sequence of the coefficients $\bar{a}_{q_i}$ is convergent. But in this case part (b) shows that $\gamma^* = \gamma$. Since $J^q(a) \to J(a)$ uniformly on K, $\lim_q J(a_q) = \lim_q J^q(a_q) = \gamma$ and (c) is proved. Part (d) is a particular case of (c) since $J(\tilde{a}) = 0$.

Remark 4.3. The above theorem remains valid if the requirements for the sets A_{ad}^M are replaced by the following. The sets $A_{\text{ad}}^M \subset A_{\text{ad}}$ are closed in $L^1(\Omega)$. For every $\hat{a} \in K$, there exists $b_M \in K \cap A_{\text{ad}}^M$ such that $b_M \to \hat{a}$ as $M \to \infty$.

5. Numerical implementation. Let Ω be a polygonal domain in $\mathbb{R}^n$ with a triangular (tetrahedral) mesh on it. Let this mesh have N internal nodes $\{p_i\}_{i=1}^N$ and M triangles

(tetrahedra) Δ_k, $1 \leqq k \leqq M$ with $\Omega = \cup_{k=1}^{M} \Delta_k$. Let $\{\phi_i\}_{i=1}^{N}$ be piecewise linear functions such that $\phi_i(p_j) = \delta_{ij}$. Define $P_M : L^1(\Omega) \to L^1(\Omega)$ by $(P_M a)(x) = 1/|\Delta_k| \int_{\Delta_k} a\, dw$ for $x \in \operatorname{int}(\Delta_k)$. Note that $\|P_M\| = 1$, $P_M A_{\text{ad}} = A_{\text{ad}}^M \subset A_{\text{ad}}$ and $P_M a \to a$ in $L^1(\Omega)$ as the diameters of the triangles obtained in mesh refinements become smaller. Let $C > 0$ and $K = \{a \in A_{\text{ad}} : \int_\Omega |Da| \leqq C\}$.

The results of the previous sections show that a coefficient $\bar{a}^q$, which solves the minimization problem

$$J^q(\bar{a}^q) = \min\{J^q(a) : a \in K \cap A_{\text{ad}}^M\}, \tag{11}$$

can be considered as an approximate solution of the parameter estimation problem (1)-(3). Here $J^q(a) = \|(u^N(a) - z)(T)\|_{L^2(\Omega)}$.

Suppose that $a \in A_{\text{ad}}^M$, that is, $a(x) = \sum_{k=1}^{M} a_k \chi(\Delta_k)$. The variation $\int_\Omega |Da|$ of $a(x)$ can be easily estimated by Theorem 2.8. (Note that all the conditions of this theorem are satisfied for $\Omega_k = \Delta_k$, $1 \leqq k \leqq M$.) Since $|\operatorname{grad} a| = 0$ on Δ_k, we find from (5) that $\int_\Omega |Da| \leqq \sum_{j=1}^{m} |a(\Gamma_j)||\Gamma_j|$ where Γ_j, $1 \leqq j \leqq m$ are the common boundaries of triangles (tetrahedra) Δ_k, $1 \leqq k \leqq M$ and $|\Gamma_j|$ are their lengths (areas). Given a coefficient $a \in A_{\text{ad}}^M$, that is, a sequence of numbers $\{a_k\}_{k=1}^{M}$, the above quantity is easily computable.

We used a gradient method to solve the problem of minimization $J^q(a)$ over the set $K \cap A_{\text{ad}}^M$. Note that the matrix $\hat{G}(a)$, defined in (6), can be represented as follows:

$$\begin{aligned}
\hat{G}(a) &= \left\{ \int_\Omega a \nabla\phi_i \nabla\phi_j \, dx \right\}_{i,j=1}^{N} \\
&= \left\{ \sum_{\Delta_k} \int_{\Delta_k} a \nabla\phi_i \nabla\phi_j \, dx \right\}_{i,j} \\
&= \left\{ \sum_{\Delta_k} \nabla\phi_i \nabla\phi_j \int_{\Delta_k} a \, dx \right\}_{i,j} \\
&= \sum_{\Delta_k} \frac{1}{|\Delta_k|} \int_{\Delta_k} a \, dx \left\{ \int_{\Delta_k} \nabla\phi_i \nabla\phi_j \, dx \right\}_{i,j} \\
&= \sum_{\Delta_k} \frac{1}{|\Delta_k|} \int_{\Delta_k} a \, dx \, G_k = \sum_{k=1}^{M} a_k G_k,
\end{aligned}$$

where $G_k = \{\int_{\Delta_k} \nabla\phi_i \nabla\phi_j \, dx\}_{i,j=1}^{N}$, and $|\Delta_k|$ is the area (volume) of Δ_k and $a_k = 1/|\Delta_k| \int_{\Delta_k} a \, dx$. Therefore (6) becomes

$$G\tilde{c}' + \sum_{k=1}^{M} a_k G_k \tilde{c} = \tilde{f}, \qquad G\tilde{c}(0) = \tilde{c}_0. \tag{12}$$

To find the gradient of $[J^q]^2 = \|u^N(a) - z\|^2$ we conclude that

$$G\left(\frac{\partial \tilde{c}}{\partial a_i}\right) + \sum_{k=1}^{M} a_k G_k \left(\frac{\partial \tilde{c}}{\partial a_i}\right) = -G_i \tilde{c}, \qquad 1 \leqq i \leqq M, \quad \frac{\partial \tilde{c}}{\partial a_i}(0) = \tilde{0},$$

and $\partial[J^q]^2/\partial a_i = 2J^q(a)\, \partial u^N/\partial a_i$.

Beginning from an initial point $a_0 \in A_{\text{ad}}$, $a_0(x) \equiv \text{const.}$ ($\int_\Omega |Da_0| = 0$), the direction of the gradient $\{\partial[J^q(a_i)]^2/\partial a_i\}_{i=1}^{M}$ is used to minimize $[J^q(a)]^2$ by direct computations of the values of $[J^q(a_0)]^2$ along this vector.

At each such point the sequence $\{a_k\}_{k=1}^{M}$ is checked to belong to the set A_{ad}^M and to satisfy the condition $\sum_{j=1}^{m} |a(\Gamma_j)||\Gamma_j| \leqq C$. The iterations of this procedure bring us to a point of minimum $\bar{a}^q$ of $J^q(a)$ over $A_{\text{ad}}^M \cap K$.

The described algorithm was implemented in Fortran and was run on the IBM-3081 at the University of Oklahoma and on the CRAY X-MP/48 at the National Center for Supercomputing Applications (University of Illinois, Urbana-Champaign). The domain Ω was taken to be $[0, 1]^2 \subset \mathbb{R}^2$ with a mesh of 105 nodes and 252 triangles for two-dimensional problems. We used $[0, 1]^3 \subset \mathbb{R}^3$ with a mesh of 729 nodes and 6,000 tetrahedra for three-dimensional problems. A typical program execution takes 40-80 min. CPU time on the IBM-3081 or 1-2 min. on the CRAY for two-dimensional problems. Three-dimensional problems require 40-80 min. CPU on the CRAY. In the later case only 1,500 (from a total of 6,000) components of grad J were found in every iteration. Other components were found from an interpolation procedure.

Here we present an example of two-dimensional identification. The test coefficient $a(x, y)$ was taken to be equal to two in the circle of radius 0.25. It is equal to one elsewhere. Figure 1 represents the graph of this test coefficient. Here the dark area corresponds to the value two and the bright area to one. By Theorem 2.8 the variation of this test coefficient is bounded by $2\pi \cdot r \cdot 1 = \pi/2$. (In fact, $\int_\Omega |Da| = \pi/2$ in this case.)

The test coefficient and the test data $z(x, y, t) = e^{-t} \sin \pi x \cdot \sin \pi y$ were substituted in (1) to obtain $f(x, t)$. This function and $u_0(x, y) = z(x, y, 0)$ were inputs of the identification program. The initial guess for $a(x)$ was $a_0(x) \equiv 1$.

Figures 2 and 3 show the identification over the sets $K_2 = \{a \in A_{ad} : \text{var}(a) \leqq 2\}$ and $K_3 = \{a \in A_{ad} : \text{var}(a) \leqq 3\}$. The difficulties in the identification in the center of the

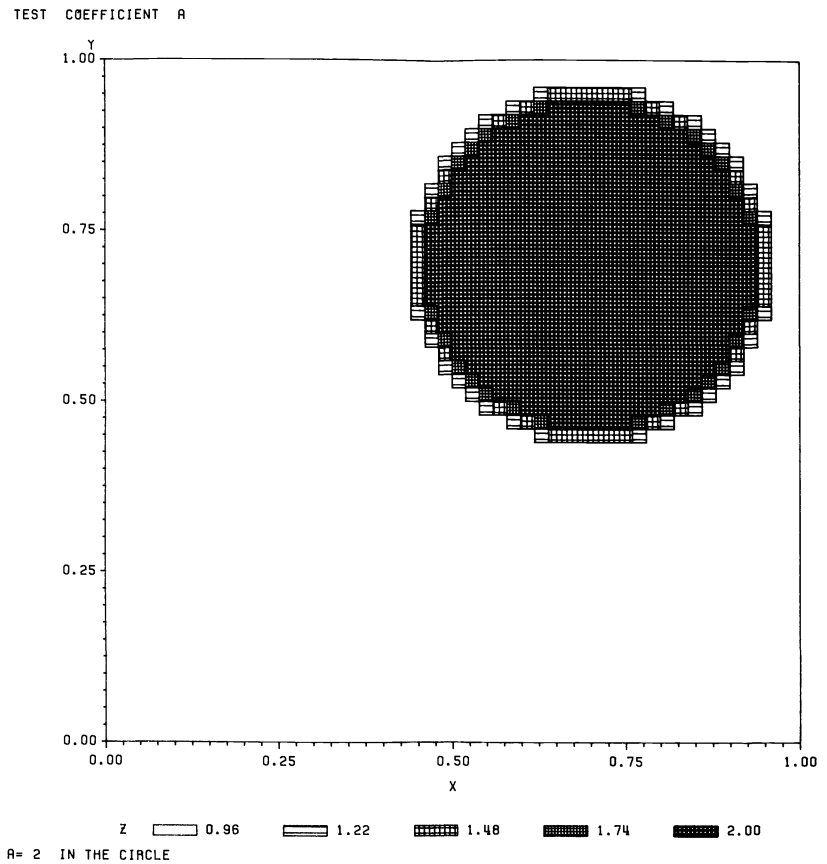


FIG. 1

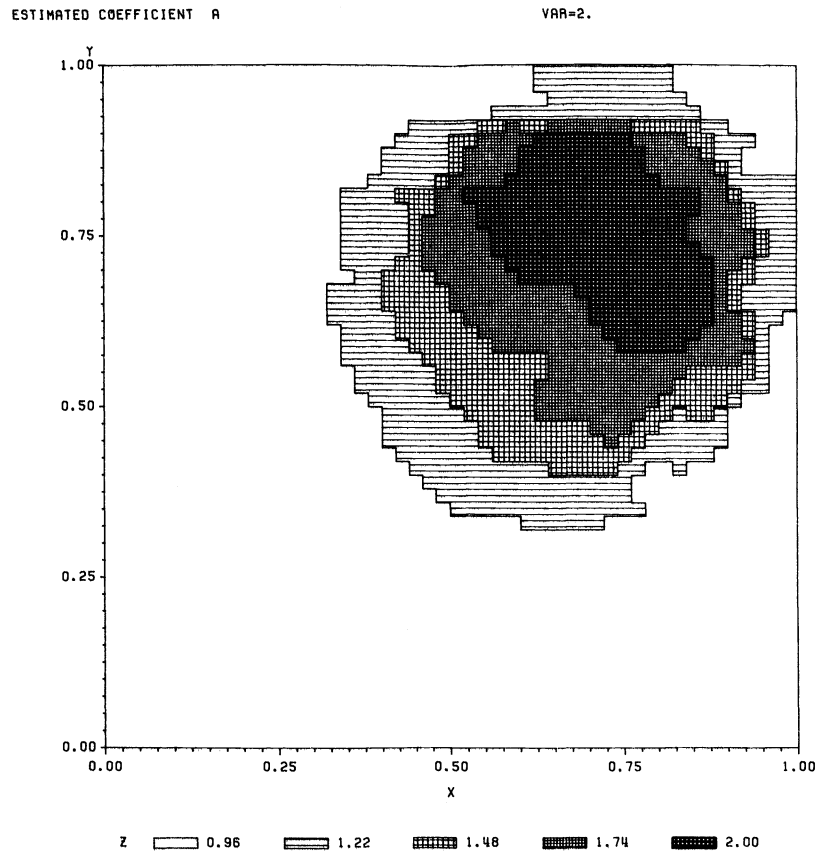


FIG. 2

domain are expected since the gradient of the test data $z(x, y, t)$ is close to zero in this region.

Figure 4 shows that the identification can be significantly improved if several sets of data are used. The above test coefficient and the test data

$$z_1(x, y, t) = e^{-t} \sin \pi x \sin \pi y,$$

$$z_2(x, y, t) = e^{-t} \sin 2\pi x \sin \pi y,$$

$$z_3(x, y, t) = e^{-t} \sin \pi x \sin 2\pi y,$$

were substituted in (1) and the functions $f_i(x, y, t)$, $i = 1, 2, 3$ were obtained. These functions and the correspondent initial conditions were considered as inputs for the identification with the cost functional $J(a) = J_1 + J_2 + J_3$, where $J_i(a) = \|(u_i^N(a) - z_i)(T)\|_{L^2(\Omega)}$. The relative $L^2(\Omega)$-error in the coefficients in the above examples is about one to two percent. Similar results were obtained for three-dimensional problems. The output (coefficients $a_{\text{test}}(x, y, z)$ and $a_{\text{est}}(x, y, z)$) was recorded on a videotape and transferred to slides. Different colors were used to indicate different values of the functions. This recording was done with the help of the Visualization Group at the National Center for Supercomputing Applications.

Appendix. We review the arguments that show $u^N \to u$ in $C([0, T]; H)$ as $N \to \infty$ (see § 3). As defined in § 3, $(\cdot, \cdot)$ is the inner product in H. We will also write (f, v),

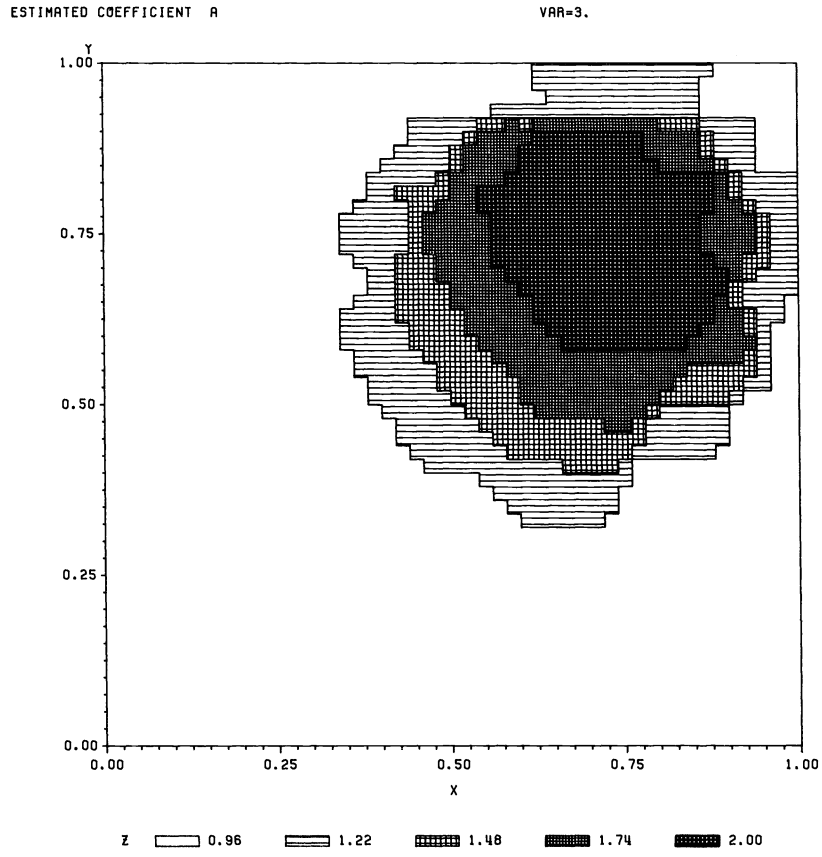


FIG. 3

$f \in V'$, $v \in V$ for the pairing between V' and V. Let $\alpha \in A_{ad}$. Define operator $A: V \to V'$ by $(Au, v) = (\alpha(x)\nabla u, \nabla v)$ for any $v \in V$. It is known [1, § 3.1.4] that $u^N \to u$ in $L^2(0, T; V)$ as $N \to \infty$. An estimate of the norm of u^N in $L^2(0, T; V)$ is given by (8). Let $V^N \subset V$ be defined as $V^N = \text{span}\{\phi_i\}_{i=1}^N$. Since V is an inner-product space and $u^N \in V^N$, we have

$$\begin{aligned}\left\| \frac{du^N}{dt} \right\|_{V'} &= \sup\left\{ \left(\frac{du^N}{dt}, w \right) : w \in V, \|w\|_V \leqq 1 \right\} \\ &= \sup\left\{ \left(\frac{du^N}{dt}, w \right) : w \in V^N, \|w\|_V \leqq 1 \right\} \\ &= \sup\{(-Au^N + f, w) : w \in V^N, \|w\|_V \leqq 1\} \\ &\leqq \mu \|u^N\|_V \|w\|_V + \|f\|_{V'} \|w\|_V,\end{aligned}$$

almost everywhere in $[0, T]$. Therefore $(\int_0^T \|du^N/dt\|_{V'}^2 \, dt)^{1/2} \leqq C$ for $N = 1, 2, \cdots$ where the constant C does not depend on N.

Similarly, $\|u_t\|_{L^2(0,T;V')} \leqq C$. Let $W(0, T) = \{y: y \in L^2(0, T; V), y_t \in L^2(0, T; V')\}$ (see, e.g., [1]). For functions $y \in W(0, T)$ we have

$$\int_0^t (y_t, y) \, dt = \frac{1}{2}|y(t)|^2 - \frac{1}{2}|y(0)|^2.$$

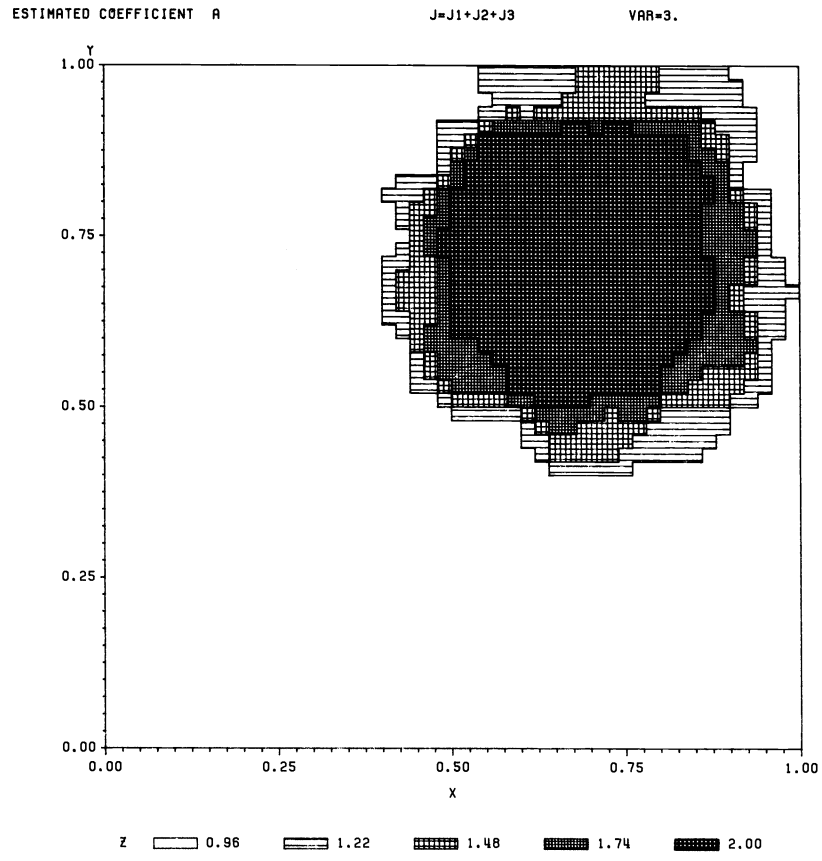


FIG. 4

Therefore

$$\int_0^t (u_t - u_t^N, u - u^N)\, dt = \frac{1}{2}|u(t) - u^N(t)|^2 - \frac{1}{2}|u(0) - u^N(0)|^2 \quad \text{and}$$

$$\|u - u^N\|^2_{C([0,T];H)} \leqq 2\int_0^T |(u_t - u_t^N, u - u^N)|\, dt + |u_0 - u^N(0)|^2.$$

By (7) $u^N(0) \to u_0$ in H as $N \to \infty$. Also

$$\begin{aligned}\int_0^T |(u_t - u_t^N, u - u^N)|\, dt &\leqq \int_0^T \|u_t - u_t^N\|_{V'} \|u - u^N\|_V\, dt \\ &\leqq \|u_t - u_t^N\|_{L^2(0,T;V')} \|u - u^N\|_{L^2(0,T;V)} \\ &\leqq 2C\|u - u^N\|_{L^2(0,T;V)},\end{aligned}$$

and the result follows.

Acknowledgments. I thank the referees for their valuable comments.

REFERENCES

[1] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[2] A. BENSOUSSAN, M. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, New York, 1978.

[3] *Distributed parameter systems: Modelling and identification*, in Proc. IFIP Working Conference, Rome 1976, A. Ruberti, ed., Lecture Notes in Control and Information Sciences, Vol. 1, Springer-Verlag, Berlin, 1978, p. 2.

[4] *Proceedings 5th* IFAC *Symposium on Identification and System Parameter Estimation*, Vol. 1, R. Isermann, ed., Pergamon Press, New York, 1980.

[5] H. T. BANKS, *On a variational approach to some parameter estimation problems*, ICASE Report No. 85-32, NASA Langley Research Center, Hampton, VA, June, 1985.

[6] C. KRAVARIS AND J. H. SEINFIELD, *Identification of parameters in distributed parameter systems by regularization*, SIAM J. Control Optim., 23 (1985), pp. 217-241.

[7] G. CHAVENT AND P. LEMONNIER, *Identification de la non-linearnité d'une equation parabolique quasilineaire*, Appl. Math. Optim., 1 (1974), pp. 121-162.

[8] K. KUNISCH AND L. WHITE, *The parameter estimation problem for parabolic equations and discontinuous observation operators*, SIAM J. Control Optim., 23 (1985), pp. 900-927.

[9] G. RICHTER, *An inverse problem for the steady state diffusion equation*, SIAM J. Appl. Math., 41 (1981), pp. 210-221.

[10] R. FALK, *Error estimates for the numerical identification of a variable coefficient*, Math. Comp., 40 (1983), pp. 537-546.

[11] V. V. ZHIKOV, S. M. KOZLOV, AND O. A. OLEINIK, *G-convergence of parabolic operators*, Russian Math. Surveys, 36 (1981), pp. 9-60.

[12] P. K. LAMM, *Estimation of discontinuous coefficients in parabolic systems: applications to reservoir simulation*, SIAM J. Control Optim., 25 (1987), pp. 18-37.

[13] S. GUTMAN AND L. W. WHITE, *On the estimation of L^∞ diffusion coefficients in parabolic equations*, preprint.

[14] S. SPAGNOLO, *Convergence of parabolic equations*, Boll. Un. Mat. Ital. (5), 14-B (1977), pp. 547-568.

[15] E. GIUSTI, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Boston, Basel, Stuttgart, 1984.

[16] P. G. CIARLET, *Numerical Analysis of the Finite Element Method*, Les Presses de L'Université de Montreal, Montreal, Quebec, Canada, 1976.

[17] R. E. EDWARDS, *Functional Analysis*, Holt, Reinhart and Winston, New York, 1965.

[18] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* I, Wiley-Interscience, New York, 1958.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 1061-1102, September 1990

003

FACTORIZATION AND THE SYNTHESIS OF OPTIMAL FEEDBACK GAINS FOR DISTRIBUTED PARAMETER SYSTEMS*

MARK H. MILMAN† AND ROBERT E. SCHEID†

Abstract. An approach based on Volterra factorization leads to a new methodology for the analysis and synthesis of the optimal feedback gain in the finite-time linear quadratic control problem for distributed parameter systems. The approach circumvents the need for solving and analyzing Riccati equations and provides a more transparent connection between the system dynamics and the optimal gain. The general results are further extended and specialized for the case where the underlying state is characterized by autonomous differential-delay dynamics. Numerical examples are given to illustrate the second-order convergence rate that is derived for an approximation scheme for the optimal feedback gain in the differential-delay problem.

Key words. factorization, Chandrasekhar equations, delay systems

AMS(MOS) subject classification. 93

1. Introduction. In this paper we develop a new synthesis methodology for the optimal feedback control law for the following general regulator problem:

$$\min_{u,x} J(u, x), \tag{1.1a}$$

$$J(u, x) = \int_0^T \langle x(t), Q(t)x(t)\rangle + |u(t)|^2 \, dt, \tag{1.1b}$$

subject to the constraint

$$x(t) = S(t)x(0) + \int_0^t S(t-r)B(r)u(r) \, dr. \tag{1.1c}$$

Here $u(\cdot)$ is H_1-valued, $x(\cdot)$ is H_2-valued (H_1 and H_2 are real separable Hilbert spaces), $Q(\cdot)$ is a strongly measurable $B(H_2)$-valued function, with $Q(t) \geqq 0$ almost everywhere, $B(\cdot)$ is a strongly measurable $B(H_1, H_2)$-valued function, and $S(\cdot)$ is a strongly continuous semigroup of operators on H_2 (cf. [9]). The general methodology is developed in Part I of the paper, while Part II of the paper specializes and extends the results to the case where (1.1) corresponds to an autonomous differential-delay system.

Solutions to systems of this type have been derived by many authors (e.g., [1], [4], [9], [23]). Although several different approaches to the problem have been developed, the feedback law is generally characterized by means of an operator Riccati equation, and numerical methods for its approximation are derived by discretizations of the infinite-dimensional Riccati equation. Gibson [9] provides a general sufficient condition for the strong operator convergence of approximating sequences of Riccati solutions. This translates in the important case in which H_1 is finite-dimensional to uniform convergence of the approximating sequence of optimal gains.

By exploiting a connection between the optimization problem (1.1) and Volterra factorization [29], we derive a new analytical characterization of the gain in terms of classical solutions of the underlying state-space dynamics. This leads to new results

* Received by the editors October 5, 1987; accepted for publication (in revised form) November 7, 1989.
† Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109.

on the differentiability properties of the gain and the development of corresponding approximation methodologies that exploit the underlying structure to guarantee the stability and accuracy of the discretizations. (Related results regarding the Riccati-based synthesis can be found in [13] and [14].)

To illustrate the fundamental connection between the optimization problem (1.1) and Volterra factorization, some motivation of the approach in the context of finite-dimensional systems will now be given. Thus we take $H_1 = H_2 = R^n$, so that $S(t) = \exp(At)$. We define the space $U = L_2(0, T; R^n)$ and the resolution of the identity E on U by $[E(\omega)u](t) = \chi(\omega)(t)u(t)$ for a Borel set $\omega \subset [0, T]$ (χ = characteristic function). The operator $S \in B(U)$ is given by

$$Su: t \to \int_0^t \exp(A(t-r))u(r)\, dr. \tag{1.2}$$

And we also define $B \in B(U)$ and $Q \in B(U)$ by the matrix definitions

$$[Bu](t) = Bu(t), \qquad [Qu](t) = Qu(t). \tag{1.3}$$

With this notation and these identifications the optimization problem (1.1) can be posed as

$$\min_u \int_0^T \langle x, Qx\rangle + |u|^2\, ds \tag{1.4a}$$

subject to the constraint

$$x = f + SBu, \qquad f(t) = \exp(At)x(0). \tag{1.4b}$$

Let $\hat{u}$ denote the solution to this problem. Next let $E_t = E(t, T)$, and let $E^t = I - E_t$, and consider for each $t \in [0, T]$ the optimization problem with objective (1.4a) but with constraint

$$x = f_t + SBE_t u, \tag{1.5}$$

where $f_t = E_t SBE^t \hat{u} + E_t f$.

The solutions $\hat{u}$ and $\hat{u}_t$ to these problems are easily obtained by completing the square and are given by

$$\hat{u} = -[I + B^*S^*QSB]^{-1}B^*S^*Qf, \tag{1.6}$$

and

$$\hat{u}_t = -[I + E_tB^*S^*QSBE_t]^{-1}E_tB^*S^*Qf_t, \tag{1.7}$$

respectively. By using a principle of optimality [29] it can be shown that $E_t\hat{u} = \hat{u}_t$. Thus if we take a partition $0 = t_0 < \cdots < t_N = T$ of $[0, T]$, using (1.7) we can write

$$\hat{u} = \sum_{i=0}^{N-1} E(\omega_i)\{I + E_{t_i}B^*S^*QSBE_{t_i}\}^{-1}B^*S^*Qf_{t_i}, \tag{1.8}$$

where $\omega_i = (t_i, t_{i+1})$. Now since $f_{t_i} = E_{t_i}SBE^{t_i}\hat{u} + E_{t_i}f$, it follows that for $\sigma \geqq t_i$,

$$\begin{aligned} f_{t_i}(\sigma) &= \exp(A\sigma)x(0) + \int_0^{t_1} \exp(A(\sigma - r))B\hat{u}(r)\, dr \\ &= \exp(A(\sigma - t_i))\left\{\exp(At_i)x(0) + \int_0^{t_i} \exp(A(t_i - r))B\hat{u}(r)\, dr\right\} \\ &= \exp(A(\sigma - t_i))\hat{x}(t_i), \end{aligned} \tag{1.9}$$

where $\hat{x}(t)$ denotes the optimal trajectory. From this we can write the last portion of the expression in (1.8) as

$$[B^*S^*Qf_{t_i}](t) = \Gamma(t, t_i)\hat{x}(t_i) \tag{1.10}$$

with

$$\Gamma(t, t_i) = B^* \int_t^T \exp(A^T(r-t))Q \exp(A(r-t_i))\, dr.$$

For each t the operator $E_tB^*S^*QSBE_t$ is Hilbert-Schmidt so that the resolvent kernel $R(t; s, \sigma)$ exists for each t, i.e.,

$$[I + E_tB^*S^*QSBE_t]^{-1} = I + R_t, \tag{1.11}$$

where R_t is the Hilbert-Schmidt operator with kernel $R(t; s, \sigma)$.

Using the expressions (1.10)-(1.11) we can return to (1.8) and write for almost every $t \in [t_i, t_{i+1}]$

$$\hat{u}(t) = \Gamma(t, t_i)\hat{x}(t_i) + R_t(\Gamma(\cdot, t_i))(t)\hat{x}(t_i). \tag{1.12}$$

Next we formally let the mesh width of the partition tend to zero and deduce that for almost every $t \in [0, T]$,

$$\hat{u}(t) = \{\Gamma(t, t) + [R_t(\Gamma(\cdot, t))](t)\}\hat{x}(t). \tag{1.13}$$

If this step could be justified, the feedback gain $K(t)$ would then have the form

$$K(t) = \Gamma(t, t) + R_t(\Gamma(\cdot, t))(t). \tag{1.14}$$

The arguments to justify this limiting procedure are developed abstractly in [29] by using the notion of Volterra factorization. In this paper we derive the "correct" version of (1.14) as well as its connection with the usual Riccati solution for the optimal gain.

This connection is particularly exploited when the state x is characterized by autonomous differential-delay dynamics (Part II). For this important special case we obtain further connections with the Krein-Bellman-Sobolev equation for close-to-displacement kernels [10], [16] and derive a new and relatively simple set of hyperbolic equations that characterizes the optimal feedback kernel. Analysis of these equations elucidates the underlying structure of the kernel and leads to the development of fast and accurate numerical methods for its computation. Unlike traditional formulations based on the operator Riccati equation, the gain is characterized by means of classical solutions of the derived set of equations.

Because of the structure of the underlying factorization problem, the corresponding algorithms for differential-delay systems are "fast" in that the operation count grows only linearly with the dimension of the discretized kernel as opposed to a quadratic or cubic growth rate for the Riccati-based synthesis [8], [15]. In fact, the equations collapse directly into the finite-dimensional Chandrasekhar equations when the delay terms are removed. However, our equations are not to be confused with the operator generalization of the Chandrasekhar equations derived in [14]. The equations we derive are for a set of quantities that possesses greater smoothness than the gain, and furthermore these equations are shown to have a classical solution (it is not an operator equation). Our new characterization of the optimal gain leads, in the case of multiple noncommensurate pure time delays with L_2 integral delay term, to a C^2 structure of the feedback kernel. For the case of a single delay with a C^∞ integral delay term, a piecewise C^∞ structure of the kernel is obtained.

In either case numerical methods can be developed to fully exploit the degree of smoothness that is available. Thus, in the general case we are able to justify methods that are second order in accuracy, and for the single-delay case described above we are able to justify methods with an arbitrarily high order of accuracy.

A brief summary of the paper follows. Part I consists of §§ 2 and 3 in which the factorization-based synthesis is developed for the system (1.1). Section 2 contains relevant background results concerning the class of operators studied in this paper. In § 3 a factorization-based representation of the optimal gain is derived, and its connection with the Riccati formalism and differentiability properties of the gain is established. Part II of the paper contains the specialization to differential-delay systems. Section 4 introduces the problem and develops a preliminary synthesis of the feedback kernel together with preliminary associated smoothness properties. In § 5 the major representation theorem for the feedback kernel for differential-delay systems is proved. In § 6 the representation theorem is used to further analyze smoothness properties of the optimal kernel. And finally, in § 7 a numerical approximation methodology is developed and implemented for two representative model problems.

Part I: General Theory

In the first part of this paper we will formalize the program outlined in the Introduction. Connections between Volterra factorizations and Riccati equations will be developed, and the differentiability properties of the relevant Volterra factors and the optimal gain will be established.

2. Notation and background results. For any Banach space Y, $|y|$ will denote the norm of an element $y \in Y$; $B(Y, Z)$ will denote the space of bounded linear maps from Y into Z, and for brevity we write $B(Y)$ for $B(Y, Y)$. Let $[0, T]$ denote a bounded closed interval in R and let Σ denote the class of Borel subsets of $[0, T]$. Now let U and X be real separable Hilbert spaces with resolutions of the identity $E_U : \Sigma \to B(U)$ and $E_X : \Sigma \to B(X)$. Assume E_U and E_X are both absolutely continuous, i.e., for each $u \in U$ and $x \in X$ the measures $|E_U(\cdot)u|^2$ and $|E_X(\cdot)x|^2$ are absolutely continuous with respect to Lebesgue measure. (In the sequel Lebesgue measure shall be denoted by λ.) A map K in $B(U, X)$ is causal if $E_X(0, t)KE_U(0, t) = E_X(0, t)K$ for all $t \in [0, T]$; K is anticausal if K^* is causal (K^* is the adjoint of K); and K is said to be memoryless if $E_X(\omega)K = E_U(\omega)K$ for all $\omega \in \Sigma$. Subscripts will be suppressed whenever the context permits. And we shall also use the notation $E^t = E(0, t)$ and $E_t = I - E^t$.

Let U and X be Hilbert spaces with resolutions of the identity E_U and E_X as defined above, and let $K \in B(U, X)$ satisfy the following hypothesis:

(H1) There exists a constant α such that
(A) $|KE(\omega)|^2 < \alpha\lambda(\omega)$, for all ω,
(B) $|E(\omega)K|^2 < \alpha\lambda(\omega)$, for all ω.

Operators satisfying (H1) are the focal point of our approach to the regulator problem, and we will devote some time in this section to developing their properties. (Details can be found in [28] and [32].) We first note the general situation in which Hilbert spaces and operators of this form arise in the paper.

Example 2.1. Let H_1 and H_2 denote real separable Hilbert spaces and let $U = L_2(0, T; H_1)$ and $X = L_2(0, T; H_2)$. Introduce the "truncation" resolutions of the identity E_U and E_X by $[E(\omega)z](t) = \chi(\omega)(t)z(t)$ for z in U or X. Let $K(t, s)$ be a strongly measurable essentially bounded $B(H_1, H_2)$-valued function and define the operator

$K \in B(U, X)$ by (the Bochner integral)

$$Ku : t \to \int_0^T K(t, s)u(s)\, ds.$$

A direct calculation shows that K satisfies (H1).

Now suppose K satisfies (H1) and define the space $H^U = L_2(0, T; U)$. Then K induces a map $F(K) \in B(H^U, X)$ in the following way.

Let $u \in H^U$ be simple, say $u(t) = \sum_{i=1}^N \chi(\omega_i)u_i$, $(u_i \in U)$. Define $F(K)u$ by the formula

$$F(K)u = \sum_{i=1}^N E(\omega_i)Ku_i.$$

Then for the simple function u we note,

$$\begin{aligned} |F(K)u|^2 &= \left|\sum_i E(\omega_i)Ku_i\right|^2 = \sum_i |E(\omega_i)Ku_i|^2 \\ &\leqq \sum_i \alpha\lambda(\omega_i)|u_i|^2 = \alpha \int_0^T |u(t)|^2\, dt = \alpha|u|^2, \end{aligned} \tag{2.1}$$

where the last norm in the chain of inequalities above is in the space H^U. Since the simple functions are dense in H^U, it follows that $F(K)$ can be extended by continuity to an operator in $B(H^U, X)$.

The operator $F(K)$ will figure prominently in subsequent analysis. The first important feature of $F(K)$ we note is that it is a memoryless map with respect to E_X and the truncation resolution E_* on H^U defined $[E_*(\omega)z](t) = \chi(\omega)(t)z(t)$.

Interpreting the resolution structure on a Hilbert space as introducing an abstract time structure, we would expect that a memoryless map should act "pointwise" between spaces in some sense. In concrete settings the next two propositions show that this is indeed the case.

PROPOSITION 2.2. *In the setting of Example* 2.1, *given any* $u \in H^U$ (H^U *defined as above*),

$$F(K)u : t \to \int_0^T K(t, s)u(t)(s)\, ds.$$

Proof. See Proposition 3.3 of [28] for the proof.

PROPOSITION 2.3. *In the setting of Example* 2.1 *suppose* $M \in B(U, X)$ *is memoryless. Then there exists a unique strongly measurable essentially bounded* $B(H_1, H_2)$*-valued function* $M(\cdot)$ *such that* $M(t)u(t) = [Mu](t)$ *almost everywhere for each* $u(\cdot) \in U$.

Proof. See Proposition 3.1 of [28] for the proof.

Note in Proposition 2.2 that if $K(t, s) = 0$ for $s < t$, then only values of $u(t)(s)$ for $s > t$ are required for computing $F(K)u$. This observation holds somewhat more generally.

PROPOSITION 2.4. *Suppose* $K \in B(U, X)$ *satisfies* (H1) *and is anticausal. Define the projection* $P^- \in B(H^U)$ *by* $[P^-u](t) = E_t u(t)$. *Then* $F(K) = F(K)P^-$.

Proof. See Lemma 4.1 of [28] for the proof.

Now given any Hilbert space Z and operator $A \in B(Z, U)$, it follows that if K satisfies (H1) then $F(KA) \in B(L_2(0, T; Z), X)$. Note that A induces an operator $\underline{A} \in B(L_2(0, T; Z), L_2(0, T; U))$ by $[\underline{A}z](t) = Az(t)$. By appealing to the action of $F(\cdot)$ on simple functions it is easy to show that

$$F(KA) = F(K)\underline{A}. \tag{2.2}$$

Proposition 2.3 and expression (2.2) are basically technical devices that will be useful in the derivation of the Riccati form of the optimal control in § 3.

PROPOSITION 2.5. *Let K satisfy* (H1), *and let H^U be defined as above with the truncation resolution E_*. Define the maps G^- and $G^+ \in B(H, H^U)$ by $[G^+u](t) = E^t u$ and $[G^-u](t) = E_t u$. Let $p_+(K) = F(K)G^+$ and $p_-(K) = F(K)G_-$. Then,*

(i) *$p_+(K)$ is causal and $p_-(K)$ is anticausal.*

(ii) *$p_+(K)$ and $p_-(K)$ satisfy* (H1).

(iii) $K = p_+(K) + p_-(K)$.

(iv) $R(p_+) \cap R(p_-) = 0$.

Proof. See Theorem 2.4 of [32] for the proof.

The proposition asserts that p_+ and p_- define projections on the vector space of maps in $B(U, X)$ that satisfy (H1). This class can be topologized so that it is a Banach space, but we will have no need to do so here.

The next three results concern invertibility and factorization within the class of operators satisfying (H1). These results form the principal components of our theory.

THEOREM 2.6. *Suppose K satisfies* (H1) *with $U = X$ and K is causal (anticausal). Then K is quasi-nilpotent and $W = (I + K)^{-1} - I$ is causal (anticausal). Furthermore, W also satisfies* (H1).

Proof. See Theorem 3.3 of [32] for the proof.

THEOREM 2.7. *Suppose K satisfies* (H1) *with $U = X$ and $I + K > 0$ is invertible. Then there exists a unique causal X satisfying* (H1) *such that*

$$(I + K) = (I + X^*)(I + X).$$

Proof. See Theorem 3.5 of [32] for the proof.

A generalization of this factorization that will be used in the Riccati synthesis of the optimal controller is given in the theorem below.

THEOREM 2.8. *Let K satisfy the hypotheses of Theorem* 2.7. *Let Y denote a Hilbert space with absolutely continuous resolution of the identity E_Y. Suppose $B \in B(Y, U)$ is memoryless. Then there exists unique causal $Z \in B(U)$ satisfying* (H1) *such that*

$$K = Z + Z^* + Z^* B B^* Z.$$

Proof. See Theorem 2.5 of [27] for the proof.

3. Representations of the optimal feedback gain. Using an abstract representation of the optimal control law developed in [29], in this section we develop two feedback representations for the optimal gain operator. The first is the standard operator Riccati formalism. We will use factorization arguments to derive the feedback gain and prove existence and uniqueness of the "first" integral Riccati equation derived by Gibson [9]. This result is not new, but the approach is quite novel and we obtain the result perhaps a bit quicker than in [9] where the "second" Riccati equation of Curtain and Pritchard [4] served as the departure point. Next, beginning again with the same abstract representation of [29], we will quickly derive an alternative feedback representation that explicitly contains the open-loop semigroup and Volterra factors in the feedback gain representation. This form of the feedback gain will be studied in greater detail in subsequent sections where imposing a finite dimensionality constraint on the input space allows us to sharpen the analysis considerably.

We consider the regulator problem and define the spaces $U = L_2(0, T; H_1)$ and $X = L_2(0, T; H_2)$ with the truncation resolutions E_U and E_X. Define the operator $S \in B(X)$ by

$$Sx : t \to \int_0^t S(t - r)x(r)\, dr, \tag{3.1}$$

and the multiplication operators $B \in B(U, X)$ and $Q \in B(X)$ by $[Bu](t) = B(t)u(t)$ and $[Qx](t) = Q(t)x(t)$.

With these identifications we can state the main theorem from [29], which is the departure point for the paper.

THEOREM 3.1. *The optimal control $\hat{u}$ for* (1.1)-(1.2) *has the representation*

$$\hat{u} = -F((I + W^*)B^*S^*Q)z(\cdot),$$

where $W = (I+X)^{-1} - I$, X is obtained from the factorization

$$(I + B^*S^*QSB) = (I + X^*)(I + X), \tag{3.2}$$

and $z(\cdot) \in C(0, T; X)$ with $z(t) = E_t S(\cdot)x(0) + E_t SBE^t\hat{u}$.

We first note that the operator B^*S^*QSB indeed satisfies the hypotheses of Theorem 2.7, so that by that theorem, Theorem 2.6, and the definition of $F(\cdot)$, the representation above makes sense. With this representation we now proceed to the Riccati formalism of the optimal gain.

THEOREM 3.2. *The optimal control $\hat{u}$ for* (1.1) *has the feedback solution $u(t) = -B^*(t)P(t)x(t)$ where $P(\cdot)$ is the unique self-adjoint solution to the integral Riccati equation*

$$P(t)x = \int_t^T S^*(r-t)\{Q(r) - P(r)B(r)B^*(r)P(r)\}S(r-t)x\, dr.$$

Proof. We begin with the representation from Theorem 3.1:

$$\hat{u} = -F((I + W^*)B^*S^*Q)z(\cdot).$$

Computing $z(t)$ explicitly we find that

$$z(t)(r) = \begin{cases} S(r)x(0) + \int_0^t S(r-\alpha)B(\alpha)\hat{u}(\alpha)\, d\alpha, & r \geqq t, \\ 0, & r < t. \end{cases}$$

Now introduce the operator $\sigma \in B(X, L_2(0, T; X))$,

$$[\sigma x](t)(r) = S(r-t)x(t).$$

Note that σ is memoryless, and furthermore since

$$\begin{aligned} z(t)(r) &= \begin{cases} S(r-t)\left[S(t)x(0) + \int_0^t S(t-\alpha)B(\alpha)\hat{u}(\alpha)\, d\alpha\right] & r > t, \\ 0, & r < t, \end{cases} \\ &= \begin{cases} S(r-t)\hat{x}(t), & r \geqq t, \\ 0, & r < t, \end{cases} \end{aligned}$$

we have $z(t) = [\sigma\hat{x}](t)$. Therefore,

$$\hat{u} = -F((I + W^*)B^*S^*Q)\sigma\hat{x}. \tag{3.3}$$

Recalling that $F(\cdot)$ and σ are both memoryless, it is easily verified that their composition is also memoryless. Hence, the operator $K = F((I + W^*)B^*S^*Q)\sigma$ is memoryless in $B(X, U)$. By Proposition 2.3 there exists an essentially bounded strongly measurable $B(H_1, H_2)$-valued function $K(\cdot)$ such that $[Kx](t) = K(t)x(t)$ almost everywhere for each $x(\cdot) \in X$. Thus,

$$\hat{u}(t) = -K(t)\hat{x}(t). \tag{3.4}$$

Now it remains to relate $K(t)$ to the Riccati equation in the statement of the theorem. We use Theorem 2.8. By the theorem we have the factorization

$$S^*QS = Z + Z^* + Z^*BB^*Z. \tag{3.5}$$

Theorem 2.6 implies Z^*BB^* is quasi-nilpotent so that we can write $Z = (I + Z^*BB^*)^{-1}(S^*QS - Z^*)$. Thus, $B^*Z = B^*(I + Z^*BB^*)^{-1}\{S^*QS - Z^*\}$. But, $B^*(I + Z^*BB^*)^{-1} = (I + B^*Z^*B)^{-1}B^*$. By uniqueness of the factorization in Theorem 2.7 (cf. (3.2) and (3.5)), we have $B^*Z^*B = X^*$. Therefore, $B^*(I + Z^*BB^*)^{-1} = (I + W^*)B^*$, and $B^*Z = (I + W^*)B^*(S^*QS - Z^*)$.

Now apply the projection p_+ to this last equality to obtain (using (2.2), Proposition 2.4, and Proposition 2.5),

$$B^*Z = F((I + W^*)B^*S^*QS)G^+ = F((I + W^*)B^*S^*Q)P^-\underline{S}G^+. \tag{3.6}$$

From the definitions of P^- (Proposition 2.4) and $\underline{S}$ (cf. (2.2)), we compute for $x \in X$,

$$[P^-\underline{S}G^+x](t) = E_tSE^tx = \begin{cases} \int_0^t S(r-\alpha)x(\alpha)\,d\alpha, & r \geqq t, \\ 0, & r < t. \end{cases}$$

But now note that

$$[\sigma Sx](t) = \begin{cases} S(r-t)\int_0^t S(t-\alpha)x(\alpha)\,d\alpha, & r \geqq t, \\ 0, & r < t. \end{cases}$$

Thus by the semigroup property of $S(\cdot)$ it follows that $P^-\underline{S}G^+ = \sigma S$, and consequently

$$B^*Z = F((I + W^*)B^*S^*Q)\sigma S = KS, \tag{3.7}$$

where K is the memoryless operator in (3.3) (hence, also (3.4)). We can argue in a similar manner to show that

$$Z = PS \tag{3.8}$$

for some memoryless operator P in $B(X)$. Since by Proposition 2.3, P can be associated with a strongly measurable essentially bounded $B(H_2)$-valued function $P(\cdot)$, it follows from (3.7) and (3.8) and standard arguments [34, p. 227] that

$$K(t)S(t-r) = B^*(t)P(t)S(t-r) \quad \text{a.e. } t, r.$$

Then from the strong continuity of $S(\cdot)$ at zero and the fact that $S(0) = I$, we obtain the identity

$$K(t) = B^*(t)P(t) \quad \text{a.e. } t. \tag{3.9}$$

To obtain the Riccati equation we substitute (3.8) into (3.5) to obtain

$$PS = S^*\{Q - P^*BB^*P\}S - S^*P^*.$$

In terms of the kernels of these maps it then follows after applying p_+ to the above that for each $x \in H_2$,

$$P(t)S(t-r)x = \int_t^T S^*(\alpha - t)\{Q(\alpha) - P^*(\alpha)B(\alpha)B^*(\alpha)P(\alpha)\}S(\alpha - t)S(t-r)x\,d\alpha.$$

Again using the strong continuity of $S(\cdot)$ and $S(0) = I$, it follows that

$$P(t)x = \int_t^T S^*(\alpha - t)\{Q(\alpha) - P^*(\alpha)B(\alpha)B^*(\alpha)P(\alpha)\}S(\alpha - t)x\,d\alpha. \tag{3.10}$$

It is evident that $P(\cdot)$ is self-adjoint, and the uniqueness of the solution to the equation above can be obtained by reversing the argument and using the uniqueness of the factorization in Theorem 2.8. (A similar argument can be found in [28].) Thus, the Riccati equation has unique solution and the theorem is completely proved. □

By using the first portion of the argument in the proof of the theorem, an alternative expression for the gain that circumvents the need for solving the operator Riccati equation will be derived. This expression, which we shall derive for the case of a finite-dimensional input space and constant B and Q, will be the focus of subsequent analysis.

So now let $H_1 = R^m$. In this case note that there exists a set of vectors $\{b_1, \cdots, b_m\} \subset H_2$ such that

$$Bu = \sum_i b_i u_i, \qquad u = (u_1, \cdots, u_m)^T \in R^m. \tag{3.11}$$

For notational convenience let $K = B^*S^*QSB$. Note that as before $K \in B(U)$, but now $U = L_2(0, T; R^m)$. Hence, since $S(\cdot)$ is a strongly continuous semigroup, and B and Q are bounded, it follows that K is a Hilbert-Schmidt operator with $m \times m$ matrix kernel $K(t, s)$, with components

$$k_{ij}(t, s) = \int_{\max(t,s)}^{T} \langle QS(r-s)b_j, S(r-t)b_i \rangle \, dr. \tag{3.12}$$

Thus the factorization of the operator $I + K$ now involves factorization of a Hilbert-Schmidt operator, and the corresponding Volterra factors are known to be Hilbert-Schmidt [11, p. 184]. Since we are now in the setting of the Lebesgue space $L_2(0, T; R^m)$, Hilbert-Schmidt operators on this space are integral operators with norm-square integrable kernel. Hence the operator W in Theorem 3.1 is itself an integral operator with $m \times m$ matrix kernel $w_{ij}(t, s)$, $i, j = 1, \cdots, m$. Furthermore, from the Riesz theorem it follows that the optimal gain $K(t)$ is also composed of a set of m vectors $(k_1(t), \cdots, k_m(t))$ in H_2. Keeping these observations in mind, we can now easily derive an alternative and useful formulation for the optimal gain.

THEOREM 3.3. *The optimal control $\hat{u}$ has the feedback form $\hat{u} = -K(t)\hat{x}(t)$ where $\hat{x}(\cdot)$ denotes the optimal trajectory and $K(t)$ denotes the optimal gain. Furthermore, in the notation above*

$$K(t)x = (\langle k_1(t), x \rangle, \cdots, \langle k_m(t), x \rangle)^T, \qquad x \in H_2,$$

where

$$k_i(t) = \int_t^T S^*(r-t)QS(r-t)b_i \, dr + \sum_{j=1}^{m} \int_t^T w_{ji}(r, t) \int_r^T S^*(\alpha - t)QS(\alpha - r)b_j \, d\alpha \, dr.$$

Proof. The proof follows from Theorem 3.1 and Proposition 2.2. Using Theorem 3.1, we have

$$\hat{u} = -F((I + W^*)B^*S^*Q)z(\cdot).$$

Next note the following representation for $(I + W^*)B^*S^*Q$. For any $x(\cdot) \in X$,

$$\begin{aligned}(I + W^*)B^*S^*Qx : t \to & \int_t^T B^*S^*(r-t)Qx(r) \, dr \\ & + \int_t^T W^*(r, t)B^* \int_{\max(t,r)}^T S^*(\alpha - r)Qx(\alpha) \, d\alpha \, dr.\end{aligned} \tag{3.13}$$

If again the operator $\sigma \in B(X, L_2(0, T; X))$ is introduced as in the proof of Theorem 3.2, and we make use of (3.3) together with the explicit forms for W^* as a Hilbert-Schmidt operator with $m \times m$ matrix kernel and the operator B as represented in (3.11), the conclusion of the theorem follows from the appropriate substitutions into (3.13) and Proposition 2.2. □

Generalizations of Theorems 3.2 and 3.3 to time-varying problems, problems with a terminal-state penalty, or problems with an infinite-dimensional input space are possible within the framework we have established. Time-varying extensions are straightforward, but a little care is required for rigorously incorporating a terminal state penalty term, although the correct representations are obtained by formally introducing delta functions into the state cost $Q(\cdot)$.

Thus if the state-space equation (1.1c) is replaced by

$$x(t) = S(t, 0)x(0) + \int_0^t S(t, r)Bu(r)\, dr,$$

where $S(t, r)$ is a strongly continuous evolution operator (see, for example, Gibson [9]) and the operator S is redefined

$$Su : t \to \int_0^t S(t, r)u(r)\, dr,$$

then the corresponding generalizations also hold. In particular, we can show that the feedback law has the form (cf. Theorem 3.3):

$$\hat{u}(t) = -K(t)\hat{x}(t),$$

$$K(t)x = (\langle k_1(t), x\rangle, \ldots, k_m(t), x\rangle)^T, \qquad x \in H_2,$$

$$k_i(t) = \int_t^T S^*(r, t)QS(r, t)b_i\, dr + \sum_{j=1}^{m} \int_t^T w_{ji}(r, t) \int_r^T S^*(\sigma, t)QS(\sigma, r)b_j\, d\sigma\, dr,$$

where as before $W(t, s)$ is the kernel of the operator W defined in Theorem 3.1. The generalization of (3.12) is given by

$$k_{ij}(t, s) = \int_{\max(t,s)}^T \langle QS(r, s)b_j, S(r, t)b_i\rangle\, dr.$$

Numerical approximations to the gain based on Theorem 3.3 will be discussed later. A matter of practical importance in this direction is the overall smoothness properties of the gain and the smoothness of the various components that comprise the expression for the gain in the theorem. The fundamental question in this regard is to what extent differentiability properties of $K(t, s)$ in (3.12) transfer to the kernel $W(t, s)$. This question is addressed in the following proposition.

PROPOSITION 3.4. *Let* $K(t, s) = K^T(s, t)$. *Suppose* $K \in C^n(\Delta)$ *where* $\Delta_- = \{(t, s): 0 < t < s < T\}$, $\Delta_+ = \{(t, s): 0 < s < t < T\}$, *and* $\Delta = \Delta_- \cup \Delta_+$. *Furthermore, assume there exists a constant* M *such that for any pair of nonnegative integers* α, β *with* $\alpha + \beta \leqq n$,

$$\sup_{0<t<s<T} |D^{\alpha,\beta}K(t, s)| \leqq M,$$

and

$$\sup_{0 \leqq t \leqq T} \left| \frac{d^\alpha}{dt^\alpha} \left\{ \lim_{s \to t} \frac{\partial^\beta K(t, s)}{\partial t^\beta} \right\} \right| \leqq M,$$

where $D^{\alpha,\beta}$ is the differential operator $\partial^{\alpha+\beta}/\partial t^{\alpha}\partial s^{\beta}$. Let K denote the integral operator on $L_2(0, T; R^m)$ with kernel $K(t, s)$ and suppose $I+K>0$. Then the Volterra factor X of $I+K$ (cf. Theorem 2.7), is an integral operator with kernel $X(t, s)$ where $X(\cdot,\cdot)\in C^n(\Delta_+)$ and satisfies the estimate

$$\sup_{t,s}|D^{\alpha,\beta}X(t, s)| = O(M)$$

for $0<s<t<T$ and $\alpha+\beta\leqq n$.

Proof. The Volterra factorization

$$I+K=(I+X^*)(I+X) \tag{3.14}$$

is equivalent to the pair of equations

$$X(t, s)=K(t, s)-\int_t^T X^*(r, t)X(r, s)\,dr, \qquad s\leqq t, \tag{3.15}$$

$$X^*(s, t)=K(t, s)-\int_s^T X^*(r, t)X(r, s)\,dr, \qquad s\geqq t. \tag{3.16}$$

Because of the relationship between the resolvent kernel of $K(t, s)$ and $X(t, s)$, $X(t, s)$ is continuous and (3.15)-(3.16) hold pointwise [11, p. 185].

Now let $W=(I+X)^{-1}-I$. W is Hilbert-Schmidt and has continuous kernel, say $W(t, s)$. Multiplying (3.14) on the left by $(I+W^*)$ and applying the projection p_+ yields the pair of identities

$$X(t, s)=K(t, s)+\int_t^T W^*(r, t)K(r, s)\,dr, \qquad s\leqq t, \tag{3.17}$$

$$X^*(s, t)=K(t, s)+\int_s^T K(t, r)W(r, s)\,dr, \qquad s\geqq t. \tag{3.18}$$

Note that $X(t, s)$ and $X^*(s, t)$ are n-times continuously differentiable in s and t, respectively. Differentiating (3.17) with respect to s and using the assumptions on $K(t, s)$, we obtain the estimate

$$\sup_{t,s}\left|\frac{\partial^n X(t, s)}{\partial s^n}\right|\leqq M(1+\delta T), \tag{3.19}$$

where $\delta=\sup|W(t, s)|$. Since $\delta=O(M)$ (see [26, Prop. 3.4]), the partial derivatives in (3.19) are also $O(M)$.

Now given a (smooth enough) function F on $[0, T]\times[0, T]$ and nonnegative integers μ and σ, we define the function F_μ^σ,

$$F_\mu^\sigma(t)=\frac{d^\sigma}{dt^\sigma}\left\{\lim_{s\to t^+}\frac{\partial^\mu F(s, t)}{\partial t^\mu}\right\}.$$

We claim that $X_\mu^{*\sigma}$ is continuous and $|X_\mu^{*\sigma}|=O(M)$ for $\mu+\sigma\leqq n$, where $X^*(s, t)$ is the function in (3.16) (equivalently in (3.18)). This assertion will be proved by induction on $p=\mu+\sigma$.

When $p=0$ the assertion is certainly true since

$$t\to K(t, t)-\int_t^T X^*(r, t)X(r, t)\,dr$$

is continuous. So assume the assertion holds for $p<n$. For any pair μ and σ with $\mu+\sigma=p+1$ (formally) calculate from (3.16)

$$X_{\mu}^{*\sigma}(t)=\frac{d^{\sigma}}{dt^{\sigma}}\left\{\lim_{s\to t^{+}}\left[\frac{\partial^{\mu}K(t,s)}{\partial t^{\mu}}-\int_{s}^{T}\frac{\partial^{\mu}X^{*}(r,t)}{\partial t^{\mu}}X(r,s)\,dr\right]\right\}$$

$$=\frac{d^{\sigma}}{dt^{\sigma}}\left\{\lim_{s\to t^{+}}\frac{\partial^{\mu}K(t,s)}{\partial t^{\mu}}\right\}-\frac{d^{\sigma}}{dt^{\sigma}}\left\{\int_{t}^{T}\frac{\partial^{\mu}X^{*}(r,t)}{\partial t^{\mu}}X(r,t)\,dr\right\}.$$

Therefore it only needs to be verified that

$$V_{\mu}(t)=\int_{t}^{T}\frac{\partial^{\mu}X^{*}(r,t)}{\partial t^{\mu}}X(r,t)\,dr$$

is σ-times continuously differentiable and $|V_{\mu}|=O(M)$ when $\sigma+\mu=p+1$.

Differentiating V_{μ} once we find

$$\dot{V}_{\mu}(t)=-X_{\mu}^{*0}X_{0}^{0}+\int_{t}^{T}\left\{\frac{\partial^{\mu+1}X^{*}(r,t)}{\partial t^{\mu+1}}X(r,t)+\frac{\partial^{\mu}X^{*}(r,t)}{\partial t^{\mu}}\frac{\partial X(r,t)}{\partial t}\right\}dr.$$

Thus, $\dot{V}_{\mu}$ involves terms with σ index equal to zero and a term of the form

$$\frac{\partial^{\mu+1}X^{*}(r,t)}{\partial t^{\mu+1}},$$

which is continuous by (3.18). Continuing in this manner until we have performed σ differentiations with $\mu+\sigma=p+1$, we find that $d^{\sigma}/dt^{\sigma}\,V_{\mu}$ involves terms of the form $X_{\alpha}^{*\beta}$ (and X_{α}^{β}) with $\alpha+\beta\leqq p$, and an additional term involving

$$\frac{\partial^{p+1}X^{*}(r,t)}{\partial t^{p+1}},$$

which again is continuous and $O(M)$ by (3.18) and (3.19). The assertion is proved.

Now return to (3.15) and (formally) calculate for $\alpha+\beta\leqq n$,

$$\frac{\partial^{\alpha+\beta}X(t,s)}{\partial t^{\alpha}\,\partial s^{\beta}}=\frac{\partial^{\alpha+\beta}K(t,s)}{\partial t^{\alpha}\,\partial s^{\beta}}-\frac{\partial^{\alpha}}{\partial t^{\alpha}}\left\{\int_{t}^{T}X^{*}(r,t)\frac{\partial^{\beta}X(r,s)}{\partial s^{\beta}}\,dr\right\}.$$

The right side above involves terms of the form

$$\frac{\partial^{r+\beta}X(t,s)}{\partial t^{r}\,\partial s^{\beta}} \tag{3.20}$$

with $r<\alpha$, and $X_{\mu}^{*\sigma}$ with $\mu+\sigma\leqq\alpha$. A straightforward induction argument now verifies that (3.20) is continuous and $O(M)$ for any nonnegative integers r and β with $r+\beta\leqq n$. □

Thus it is seen that no derivatives are lost in the factorization problem. It is not difficult to verify that the resolvent kernel $W(t,s)$,

$$W(t,s)=-X(t,s)-\int_{s}^{t}X(t,r)W(r,s)\,dr$$

enjoys the same differentiability properties as $X(t,s)$.

For future development it is also necessary to establish stability of the factorization problem above. The following corollary to Proposition 3.4 essentially shows that the factorization problem is "well-posed" in the space of integral operators with kernels in $C^{n}(\Delta)$, where $\Delta=\Delta_{+}\cup\Delta_{-}$ (cf. notation in Proposition 3.4). This result is analogous to the one in [26].

COROLLARY 3.5. *Let K, X, W denote the operators in Proposition* 3.4. *Let D be a self-adjoint integral operator with kernel $D(t, s)$ and assume $D(\cdot, \cdot) \in C^n(\Delta)$. In addition suppose that for any pair of nonnegative integers α and β with $\alpha + \beta \leqq n$ that*

$$\sup_{0<t<s<T} \left| \frac{\partial^{\alpha+\beta}}{\partial t^\alpha \, \partial s^\beta} D(t, s) \right| = \delta.$$

Then if δ is sufficiently small, $I + K + D$ has the Volterra factorization

$$I + K + D = (I + X_{D^*})(I + X_D),$$

and furthermore,

$$\sup_{0<s<t<T} \left| \frac{\partial^{\alpha+\beta}}{\partial t^\alpha \, \partial s^\beta} \{X_D(t, s) - X(t, s)\} \right| = O(\delta).$$

Proof. The proof of this result is basically the same as in the analogous proposition in [26]; thus we will only sketch the argument.

Since $|D| \to 0$ as $\delta \to 0$, it follows from Theorem 2.7 that for δ sufficiently small $I + (I + W^*)D(I + W)$ has a Volterra factorization

$$I + (I + W^*)D(I + W) = (I + Z_D^*)(I + Z_D)$$

with Z_D causal. But note that

$$\begin{aligned} I + K + D &= (I + X^*)(I + X) + D \\ &= (I + X^*)[I + (I + W^*)D(I + W)](I + X) \\ &= (I + X^*)(I + Z_D^*)(I + Z_D)(I + X). \end{aligned}$$

Hence, $X_D = X + Z_D + Z_D X$, since

$$I + K + D = (I + X_D^*)(I + X_D).$$

But,

$$X_D - X = Z_D(I + X).$$

The result then follows from Proposition 3.4. □

With Proposition 3.4 it is possible to prescribe some regularity conditions of the optimal gain via the representation in Theorem 3.3 and hypotheses concerning the open-loop semigroup $S(\cdot)$, the operator Q, and the vectors $\{b_i\}$. However, even in the general setting of the control problem where only continuity of the kernel $K(t, s)$ can be assumed (since no further assumptions are imposed on the semigroup $S(t)$, etc.); it is still possible to retrieve differentiability of the optimal gain. This is proved in the theorem below.

THEOREM 3.6. *Let $k_i(t)$ be defined as in Theorem* 3.3. *Then $t \to k_i(t)$ is strongly differentiable.*

Proof. For notational simplicity we will consider the scalar input case, i.e., $m = 1$. We introduce the function

$$v(t) = \int_0^t S^*(r) Q S(r) b \, dr.$$

(Here $b = b_1 = b_m$.) Then we can write

$$k(t) = v(T - t) + \int_t^T W(r, t) S^*(r - t) v(T - r) \, dr.$$

Making the change of variables $\mu \to r - t$ in the integrand above, we obtain

$$k(t) = v(T-t) + \int_0^{T-t} W(t+\mu, t) S^*(\mu) v(T-\mu-t)\, d\mu.$$

Since $v(t)$ is strongly differentiable, it suffices to prove that $t \to W(t+\mu, t)$ is differentiable. To this end define

$$f(t) = Q^{1/2} S(T-t) b,$$

and let $f_i(t) = \langle f(t), e_i \rangle e_i$ where $\{e_i\}$ is a complete orthonormal basis for H_2. Next let

$$F_j(t) = \sum_{i=1}^{j} f_i(t).$$

Note that Dini's theorem implies $F_j(t) \to f(t)$ uniformly since

$$\sum_{i=j}^{\infty} |f_i(t)|^2 = \sum_{i=j}^{\infty} |\langle Q^{1/2} S(T-t) b, e_i \rangle|^2$$

shows that for each t, $|F_j(t) - f(t)|$ converges monotonically to zero.

Associate $F_j(t)$ with the R^j-valued function $(\langle f(t), e_1 \rangle, \cdots, \langle f(t), e_j \rangle)^T$ and denote this function again by $F_j(t)$. Let K_j denote the integral operator with kernel

$$K_j(t, s) = \int_{\max(t,s)}^{T} F_j^T(r-t) F_j(r-s)\, dr.$$

Then clearly $K_j(t, s) \to K(t, s)$ where $K(t, s)$ is defined in (3.12). Thus [26] shows that $W_j(t, s) \to W(t, s)$ uniformly, where $W_j(t, s)$ denotes the corresponding Volterra factor associated with $K_j(t, s)$.

Now from [16] we have the following set of equations: For $t \leqq s \leqq T$

$$q_j(t) = F_j^T(t) + \int_t^T W_j(r, t) F_j^T(r)\, dr,$$

$$q_j(t, s) = q_j(s) + \int_t^s W(s, r) q_j(r)\, dr,$$

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right) W_j(s, t) = q_j(t) q_j^T(t, s), \qquad W_j(T, t) = 0.$$

Setting $s = t + \mu$ and using the boundary conditions, we obtain

$$\begin{aligned} W_j(t+\mu, t) &= -\int_t^{T-\mu} q_j(r) q_j^T(r, r+\mu)\, dr \\ &= -\int_t^{T-\mu} q_j(r) \left\{ q_j^T(r+\mu) + \int_r^{\mu+r} W_j(r+\mu, \alpha) q^T(\alpha)\, d\alpha \right\} dr. \end{aligned}$$

Since $W_j(r+\mu, \alpha) \to W(r+\mu, \alpha)$ uniformly (with $W(\cdot, \cdot)$ continuous), it suffices to show that

$$q_j(\sigma) q_j^T(r) \to p(\sigma, r),$$

uniformly to a continuous function $p(\sigma, r)$. For then by dominated convergence we get the relationship

$$W(t+\mu, t) = -\int_t^{T-\mu} \left[p(\sigma, \sigma+\mu) + \int_\sigma^{\sigma+\mu} p(\sigma, r) W(\sigma+\mu, r)\, dr \right] d\sigma,$$

from which it is clear that $t \to W(t+\mu, t)$ is differentiable.

So first observe that by the triangle inequality

$$|q_j(\sigma)q_j^T(r)-q_k(\sigma)q_k^T(r)|\leqq|q_j^T(r)||q_j(\sigma)-q_k(\sigma)|+|q_k(\sigma)||q_j^T(r)-q_k^T(r)|. \tag{3.21}$$

Since $F_j(\cdot)$ and $W_j(\cdot,\cdot)$ are both uniformly bounded, it follows that

$$\sup_j \sup_t |q_j(t)|\leqq c, \tag{3.22}$$

for some constant c. Furthermore,

$$|q_j(t)-q_k(t)|\leqq\int_t^T |W_j(\sigma,t)F_j^T(\sigma)-W_j(\sigma,t)F_k^T(\sigma)|\,d\sigma + \int_t^T |W_j(\sigma,t)F_k^T(\sigma)-W_k(\sigma,t)F_k^T(\sigma)|\,d\sigma.$$

Then since $W_j(\cdot,\cdot)\to W(\cdot,\cdot)$ uniformly and $F_k(\cdot)\to f(\cdot)$ uniformly,

$$|q_j(t)-q_k(t)|\to 0 \quad \text{uniformly.}$$

Hence, by (3.21)-(3.22), $|q_j(\sigma)q_j^T(r)-q_k(\sigma)q_k^T(r)|\to 0$ uniformly on $[0,T]\times[0,T]$ as $j,k\to\infty$, and the theorem is proved. □

This result was proved in an entirely different manner in [14] using some recently derived differentiability properties of solutions to operator Riccati equations.

The constructive nature of the techniques we have developed leads readily to the derivation of stable and accurate difference approximations for the gain based on the formulation given in Theorem 3.3. The approximation process essentially can be decoupled into three stages. First we require the approximate solution of the state equations for the space of control inputs ($i\in\{1,\cdots,m\}$):

$$\begin{aligned} &S(t)b_i, \qquad 0\leqq t\leqq T,\\ &S^*(s)QS(t)b_i, \qquad 0\leqq t\leqq T,\quad 0\leqq s\leqq T-t. \end{aligned} \tag{3.23}$$

Second, we require the approximate solution of the factorization problem (3.2) (see, for example, [10], [25]). And finally we require approximations for the quadratures given in Theorem 3.3.

This justifies an approximation methodology for the combined problem since each component of the process can be treated by well-known methods that are stable and accurate. More details are given in [31].

Part II: Application to Differential Delay Systems

In this part of the paper we will show how the general theory of Part I can be applied to the particular problem of synthesizing control gains for the optimal regulator problem with differential-delay dynamics. We will see that the delay problem is very amenable to these methods and provides a very convenient framework for exploiting connections between factorization problems, Riccati equations, and "fast" algorithms for the solution of the Krein-Bellman-Sobolev equation [10].

4. Two solutions for the optimal gain. For the remainder of this paper we will be concerned with the following regulator problem with dynamics

$$\begin{aligned} &\dot{x}(t)=\sum_{i=0}^{\nu} A_i x(t-r_i)+\int_{-r}^{0} A(s)x(t+s)\,ds+Bu(t), \qquad t>0,\\ &x(t)=\phi(t), \qquad t\in[-r,0], \end{aligned} \tag{4.1}$$

and quadratic cost criterion

$$J(u, x) = \int_0^T \langle x(t), Qx(t)\rangle + |u(t)|^2\, dt. \tag{4.2}$$

Here we assume $x(t) \in R^n$, $u(t) \in R^m$, and $\phi(\cdot) \in L_2(-r, 0; R^n)$, and $0 = r_0 < \cdots < r_\nu$. We also assume that A_i and B are matrices of compatible dimensions, $A(\cdot) \in L_2(-r, 0; R^{n\times n})$ and $Q \geqq 0$. Without loss of generality we take $r = r_\nu$.

We introduce the state space $M_2 = R^n \times L_2(-r, 0; R^n)$ with the canonical inner product and projection Π: $M_2 \to M_2$ defined by

$$\Pi(x^0, x(\cdot)) = (x^0, 0).$$

We also define the maps

$$\tilde{B}\colon R^m \to M_2, \qquad \tilde{B}v = (Bv, 0),$$

$$\tilde{Q}\colon M_2 \to M_2, \qquad \tilde{Q}(x^0, x(\cdot)) = (Qx^0, 0).$$

Then (4.1)-(4.2) is equivalent to the M_2 state-space regulator problem

$$\dot{x} = Ax + \tilde{B}u, \qquad x(0) = (\phi(0), \phi(\cdot)) \in M_2, \tag{4.1'}$$

$$J(u, x) = \int_0^T \langle x(t), \tilde{Q}x(t)\rangle_{M_2} + |u(t)|^2\, dt, \tag{4.2'}$$

where A: $D(A) \to M_2$, $D(A) = \{(x, \phi)\colon \phi' \in L_2(-r, 0; R^n), x = \phi(0)\}$, and

$$A((\phi(0), \phi(\cdot))) = \left(\sum_{i=0}^{p} A_i\phi(-r_i) + \int_{-r}^{0} A(s)\phi(s)\, ds, \phi'\right). \tag{4.3}$$

A generates a C_0 semigroup of operators $\{S(t)\}$ on M_2 (see, for example, [5]), so that (4.1') has solution

$$x(t) = S(t)x(0) + \int_0^t S(t-r)Bu(r)\, dr, \tag{4.4}$$

and Theorem 3.3 is immediately applicable.

Let $Y(t)$ denote the fundamental matrix solution for (4.1), i.e., Y satisfies the homogeneous equation with initial condition $Y(0) = I$ (the $n \times n$ identity matrix). Then the relevant matrix kernel $K(t, s)$ (cf. (3.12)) is easily shown to have the form

$$K(t, s) = \int_{\max(t,s)}^{T} B^* Y^*(r-t)QY(r-s)B\, dr. \tag{4.5}$$

Let $W(t, s)$ denote the associated factorized kernel as in Theorem 3.3. We obtain the following characterization of the gain.

THEOREM 4.1. *Let the functions W and Y be defined as above. Then the optimal control $\hat{u}$ for* (4.1)-(4.2) *has the feedback form*

$$\hat{u}(t) = -P(t, t)\hat{x}(t) - \int_t^{\min(t+r,T)} P(t, \alpha) \int_{t-r}^{t} A(s-\alpha)\hat{x}(s)\, ds\, d\alpha$$

$$- \sum_{i=1}^{\nu} \int_t^{\min(t+r_i,T)} P(t, \alpha)A_i\hat{x}(\alpha - r_i)\, d\alpha,$$

where the $m \times n$ matrix kernel $P(t, \alpha)$ is given by

$$P(t, \alpha) = L(t, \alpha) + \int_t^T W^*(s, t)L(\alpha, s)\, ds$$

with

$$L(t,\alpha)=\int_{\alpha}^{T} B^{*}Y^{*}(s-\alpha)QY(s-t)\,ds.$$

Furthermore, $P(t,\alpha)$ is C^1 on $\alpha>t$ with bounded (Fréchet) derivative.

Proof. Given $t\geqq 0$, let $x_t(\cdot)$ denote the translated function $x_t(s)=x(t+s)$, $s\in[-r,0]$. Now let $x\in M_2$, say $x=(x^0,x(\cdot))$. Note that $\Pi S(\mu-t)x$ solves the equation

$$\frac{dz}{d\mu}=\sum_{i=0}^{\nu}A_i z(\mu-r_i)+\int_{-r}^{0}A(s)z(\mu+s)\,ds$$

with initial condition

$$z(\mu)=\begin{cases}x^0, & \mu=t,\\ x(s), & \mu=t+s,\quad s\in[-r,0).\end{cases}$$

In particular, by the variation of constants formula [12, p. 148],

$$\begin{aligned}\Pi S(\mu-t)x=\;&Y(\mu-t)z(t)+\sum_{i=1}^{\nu}\int_{t-r_i}^{t}Y(\mu-\alpha-r_i)A_i z(\alpha)\,d\alpha\\&+\int_{t}^{t+r}Y(\mu-s)\left\{\int_{t-r}^{t}A(\alpha-s)z(\alpha)\,d\alpha\right\}ds.\end{aligned}\tag{4.6}$$

For notational convenience, in what follows we will consider the single input case in which $m=1$. Thus, the matrix B of (4.1) now consists of the single vector b. With this notation define the function $L(\cdot,\cdot)$ by

$$L^{T}(t,\alpha)=\int_{\alpha}^{T}Y^{T}(r-t)QY(r-\alpha)b\,dr.$$

From Theorem 3.3 the optimal gain $K(t)$ is characterized via

$$\begin{aligned}\langle K(t),x\rangle=&\int_{t}^{T}\langle QS(r-t)b,\Pi S(r-t)x\rangle\,dr\\&+\int_{t}^{T}W^{*}(r,t)\int_{r}^{T}\langle QS(\mu-r)b,\Pi S(\mu-t)x\rangle\,d\mu\,dr.\end{aligned}$$

Noting (4.6), an interchange in the order of integration above using the function $L(t,\alpha)$ yields

$$\begin{aligned}\langle K(t),x\rangle=&\langle L^{T}(t,t),x^0\rangle+\sum_{i=1}^{\nu}\int_{t}^{t+r_i}\langle L^{T}(t,\alpha),A_i x(\alpha-r_i)\rangle\,d\alpha\\&+\int_{t}^{t+r}\left\langle L^{T}(t,\alpha),\int_{t-r}^{t}A(s-\alpha)x(s)\,ds\right\rangle d\alpha\\&+\int_{t}^{T}\langle W(r,t)L^{T}(t,r),x^0\rangle\,dr\\&+\sum_{i=1}^{\nu}\int_{t}^{t+r_i}\left\langle\int_{t}^{T}W(r,t)L^{T}(\alpha,r)\,dr,A_i x(\alpha-r_i)\right\rangle d\alpha\\&+\int_{t}^{t+r}\int_{\alpha}^{T}\left\langle W(r,t)L^{T}(\alpha,r)\,dr,\int_{t-r}^{t}A(s-\alpha)x(s)\,ds\right\rangle d\alpha,\end{aligned}$$

where $x=(x^0, x_t)$. Thus if we introduce $P(t,\alpha)\in R^{1\times n}$

$$P(t,\alpha)=L^T(t,\alpha)+\int_\alpha^T W(r,t)L^T(\alpha,r)\,dr,$$

the expression above may be written more compactly as

$$\langle K(t),x\rangle=\langle P^T(t,t),x\rangle+\sum_{i=1}^{\nu}\int_t^{t+r_i}\langle P^T(t,\alpha),A_ix(\alpha-r_i)\rangle\,d\alpha$$
$$+\int_t^{t+r}\left\langle P^T(t,\alpha),\int_{t-r}^{t}A(s-\alpha)x(s)\,ds\right\rangle d\alpha.$$

This is precisely the representation of the theorem for the scalar (i.e., $m=1$) input case.

Using the fact that $Y(\cdot)$ is absolutely continuous, it follows that L is C^1 on $\alpha>t$ with bounded derivative, and also that $K(t,s)$ (cf. (4.5)) satisfies Proposition 3.4. Thus, that P is C^1 on $\alpha>t$ with bounded derivative is immediate from its definition. The extension to the multi-input case with $m>1$ is straightforward, and is omitted. The theorem is completely proved. □

The Riccati formalism of Theorem 3.2 leads to a representation of the optimal gain $K(t)$ as an element of $B(M_2, R^m)$. The gain in this representation can be realized in component form as [8]

$$K(t)=(K^{00}(t), K^{01}(t,\cdot)), \tag{4.7}$$

so that for a vector $\tilde{x}=(x^0, x(\cdot))\in M_2$,

$$K(t)\tilde{x}=K^{00}(t)x^0+\int_{-r}^{0}K^{01}(t,\alpha)x(\alpha)\,d\alpha.$$

Putting this description of the gain together with Theorem 4.1 yields the identities

$$K^{00}(t)=P(t,t), \tag{4.8}$$

$$K^{01}(t,\alpha)=\sum_{i=1}^{\nu}\chi[-r_i,0](\alpha)P(t,\alpha+t+r_i)A_i+\int_t^{t+r}P(t,s)A(\alpha+t-s)\,ds. \tag{4.9}$$

In addition to these identities, there is one other fundamental background connection between the factorization and Riccati based approaches that will be used in the next section. This is contained in the following theorem.

THEOREM 4.2. *Let $X(t,s)$ solve the resolvent identity*

$$X(t,s)+W(t,s)+\int_s^t X(t,r)W(r,s)\,dr=0,$$

and let $K(t)$ denote the optimal M_2 state space gain as above. Then

$$X(t,s)=K(t)S(t-s)\tilde{B}.$$

Proof. See Theorem 4.2 of [28] for the proof.

5. Representation theorem. In this section we will combine the representations of the previous section to derive a new set of equations for the optimal gain. The first step toward combining the gain representations is the following observation.

PROPOSITION 5.1. $-P(t,\alpha)B=W^T(\alpha,t)$ *for all* $\alpha\geqq t$.

Proof. Let P denote the Hilbert-Schmidt operator with kernel $P(t,\alpha)$, and let Y denote the operator with kernel $Y(t-\alpha)$. For a Hilbert-Schmidt operator M, we will

also write $[M]_-$ for $p_-(M)$ (cf. Proposition 2.5). In [26] it is shown that P uniquely solves the Wiener-Hopf-type equation

$$P=[B^*Y^*QY]_--[PBB^*QY]_-, \tag{5.1}$$

where B (respectively, B^* and Q) are interpreted as operators; $Bu: t\to Bu(t)(B^*x: t\to B^Tx(t), Qx: t\to Qx(t))$. Let K denote the Hilbert-Schmidt operator with kernel $K(t, s)$ given in (4.5). Multiplying the equation above on the right by the (operator) B we obtain the identity

$$PB=[K]_--[PBK]_-, \tag{5.2}$$

since $B^*Y^*QYB=K$.

In a manner similar to that in [26] we can show that PB is the unique Hilbert-Schmidt operator that satisfies (5.2). However, note that

$$\begin{aligned} W^*+[W^*K]_- &= [W^*(I+K)]_- \\ &= [W^*(I+W^*)^{-1}(I+X)]_- \\ &= [-X^*-X^*X]_-. \end{aligned}$$

And from Theorem 2.7, $K=X+X^*+X^*X$. Hence, $[K]_-=[X^*+X^*X]_-$ since $p_-(X)=0$. It then follows that $-W^*$ also solves (5.2), and so by uniqueness $W^*=-PB$. Consequently, $W^T(\alpha, t)=-P(t, \alpha)B$ almost everywhere on $[0, T]\times[0, T]$ (see, for example, [34, p. 227]). But $W(\cdot, \cdot)$ and $P(\cdot, \cdot)$ are both continuous on $\{(\alpha, t): \alpha \geqq t\}$, and the proposition is proved. □

Thus we see that the relationship between the feedback kernel $P(\cdot, \cdot)$ and the factorized kernel $W(\cdot, \cdot)$ is much more direct than was indicated in Theorem 4.1. The next theorem very nearly provides a set of differential equations with which to compute $W(\cdot, \cdot)$.

THEOREM 5.2 ([10]; also [16]). *Let* $f(t)=Q^{1/2}Y(T-t)B$ *and define* $\phi(\cdot)$

$$\phi(t)=f^T(t)+\int_t^T W^T(\sigma, t)f^T(\sigma)\,d\sigma.$$

Then for $s\geqq t$,

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial s}\right)W^T(s, t)=\phi(t)\phi^T(t, s), \qquad W(T, t)=0,$$

$$\frac{\partial}{\partial t}\phi(t, s)=-W(s, t)\phi(t), \qquad \phi(t, t)=\phi(t).$$

These equations arise in the solution to the Krein-Bellman-Sobolev equations for close to displacement kernels [10], [16]. In [10] it is shown that $W(t, s)$ can be computed on an $N\times N$ grid in $O(N^2)$ operations with $O(1/N)$ accuracy. These equations are the prime motivation for the main representation theorem of the paper, which is to follow. The proof is quite long, so we will provide a brief outline before beginning.

Assume that $B\in R^{n\times n}$ is invertible. Then Proposition 5.1 allows us to write $P(t, \alpha)=-W^T(\alpha, t)B^{-1}$. If we consider the relations in Theorem 5.2, multiplying the equation for W^T on the right by B^{-1} and the equation for ϕ on the left by B^{-T}, it follows that

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial s}\right)P^T(t, s)=\phi(t)[\phi^T(t, s)B^{-1}],$$

$$\frac{\partial}{\partial t}[B^{-T}\phi(t, s)]=P^T(t, s)\phi(t).$$

The main bottleneck to overcome in this set of equations is having to solve for $\phi(t)$, which in turn requires $W(\sigma, t)$ on the set $\{(\sigma, t): t \leqq \sigma \leqq T\}$. Note that $P(t, s)$ itself is only required for $s \in (t, t+r)$. A substantial savings in computation can be afforded by finding an alternative method for computing $\phi(t)$.

In the development of these equations, we have yet to exploit anything of the underlying state-space characteristics of the problem. In the theorem we do precisely that by using the relationship between $W(t, s)$ and the optimal gain (via the resolvent kernel $X(t, s)$ and Theorem 4.2). In this way a differential equation is developed for ϕ. Perturbation arguments, using the well posedness of the factorization and control problem, allow us to remove the invertibility condition on B.

Without loss of generality we assume that $B \in R^{n\times n}$ below (if $m < n$ we augment B with $n-m$ columns of zeros; if $m > n$, a change of coordinates in R^m reduces B to a matrix with $m-n$ zero columns, which can be discarded before defining the problem).

THEOREM 5.3. *The following set of equations*

$$
\begin{aligned}
&\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial s}\right)P(t, s) = -B^T\psi^0(t)\phi^T(t, s), \qquad 0<t<s<T,\\
&\frac{\partial}{\partial t}\phi(t, s) = P^T(t, s)B^T\psi^0(t), \qquad 0<t<s<T,\\
&\frac{d}{dt}\psi^0(t) = -[A_0^T - P^T(t, t)B^T]\psi^0(t) - \sum_{i=1}^{\nu} A_i^T\phi(t, t+r_i)\\
&\qquad\qquad -\int_{-r}^{0} A^T(\sigma)\phi(t, t-\sigma)\,d\sigma, \qquad 0<t<T,
\end{aligned}
\tag{5.3}
$$

with boundary conditions

$$
\begin{aligned}
&P(t, T) = 0,\\
&\phi(s, s) = \psi^0(s), \quad \phi(t, s) = 0, \quad s > T,\\
&\psi^0(T) = Q^{1/2}, \quad \psi^0(s) = 0, \quad s > T,
\end{aligned}
\tag{5.4}
$$

has a unique solution in the class of continuous functions on the closure of their respective domains. P is C^1 with a uniformly continuous derivative on $0<t<s<T$, ψ^0 is piecewise C^1 with a bounded derivative on $0<t<T$, and $\phi(\cdot, s)$ is C^1 for each s and $\phi(t, \cdot)$ is piecewise C^1 with a bounded derivative for each t on $0<t<s<T$. Furthermore, $P(t, s)$ is the optimal feedback kernel for (4.1)-(4.2).

Proof. We begin by assuming that the matrix B is invertible. In operator notation the equation for ϕ in Theorem 5.2 is written

$$\phi = f^T + W^* f^T. \tag{5.5}$$

Using $X^* = (I + W^*)^{-1}$, this is equivalent to

$$\phi + X^*\phi = f^T. \tag{5.6}$$

Observe from Theorem 4.2 that the kernel of X^*, say $X^*(t, \sigma)$, is

$$X^*(t, \sigma) = \tilde{B}^* S^*(\sigma - t)K^*(\sigma),$$

where $K(\cdot)$ denotes the optimal (M_2 state space) gain for the control problem. Thus (5.6) leads to

$$\phi(t) + \int_t^T \tilde{B}^* S^*(\sigma - t)K^*(\sigma)\phi(\sigma)\,d\sigma = f^T(t). \tag{5.7}$$

Let $\{e_i\}$, $i=1, \cdots, n$ denote the standard orthonormal basis for R^n, and note that $\tilde{B}^*S^*(T-t)\tilde{Q}^{1/2}$ can be identified with $f^T(t)$ via $\tilde{B}^*S^*(T-t)\tilde{Q}^{1/2}(e_j, 0)=f^T(t)e_j$ (the jth column of f^T). Let $\phi_j(t)$ denote the jth column of $\phi(t)$ and introduce the M_2-valued functions $h_j(t)$ and $\psi_j(t)$:

$$h_j(t)=-\int_t^T S^*(\sigma-t)K^*(\sigma)\phi_j(\sigma)\,d\sigma, \tag{5.8}$$

$$\psi_j(t)=h_j(t)+S^*(T-t)\tilde{Q}^{1/2}(e_j, 0). \tag{5.9}$$

Multiplying the equation above on the left by $\tilde{B}^*$ and comparing it with (the jth column of) (5.7), we arrive at the identity

$$\phi_j(t)=\tilde{B}^*\psi_j(t). \tag{5.10}$$

Let A^* denote the infinitesimal generator of S^*. (Hence A^* is the adjoint of A.) We note that [9]

$$\begin{aligned} D(A^*)=\{&(\alpha, x(\cdot))\in M_2 : x \text{ is absolutely continuous except at the points } \{r_i\},\\ &\text{where } x(r_i)^+-x(r_i)^-=A_i^T\alpha;\ x'(\cdot)\in L_2(-r_{i+1}, -r_i; R^n),\\ &i=0, \cdots, \nu-1; \text{ and } x(-r_\nu)=A_\nu^T\alpha\}, \text{ and} \end{aligned}$$

$$A^*(\alpha, x(\cdot))=(A_0^T\alpha+x(0), A^T(\cdot)x(\cdot)-x'(\cdot)).$$

We now make the following two claims.

CLAIM 1. *$h_j(\cdot)$ is differentiable, $h_j(t)\in D(A^*)$ for $t\leqq T$, and pointwise satisfies the equation*

$$\frac{d}{dt}h_j(t)=-A^*h_j(t)+K^*(t)\phi_j(t), \qquad h_j(T)=0. \tag{5.11}$$

CLAIM 2. *Let $\psi_j^0(t)$ denote the R^n component of $\psi_j(t)$. Define the R^n-valued function $\phi_j(t, s)$ by*

$$\phi_j(t, s)=\phi_j(s)+\int_t^s W(s, \sigma)\phi_j(\sigma)\,d\sigma, \qquad s\geqq t.$$

(We note that $\phi_j(\cdot, \cdot)$ is the jth column of $\phi(\cdot, \cdot)$ in Theorem 5.2.) Extend $\phi_j(\cdot)$ (and, by definition $\phi_j(t, \cdot)$) to (T, ∞) by $\phi_j(s)=0$ for $s>T$. Define $\tilde{\phi}_j(t, \alpha)$ on $[0, T]\times[-r, 0]$ by

$$\tilde{\phi}_j(t, \alpha)=\sum_{i=1}^{\nu}\phi_j^T(t, t+r_i+\alpha)\chi[-r_i, 0](\alpha)B^{-1}A_i+\int_0^r \phi_j^T(t, s+t)B^{-1}A(\alpha-s)\,ds.$$

We claim that

$$h_j(t)=(\psi_j^0(t), \tilde{\phi}_j^T(t, \cdot))-S^*(T-t)\tilde{Q}^{1/2}(e_j, 0). \tag{5.12}$$

These claims will be verified at the end of the proof of the theorem, and will be accepted for now.

Write $h_j(t)$ in component form as $h_j(t)=(h_j^0(t), h_j^1(t))$. From (5.11) and the definition of A^* above, the following equation holds for h_j^0:

$$\begin{aligned} &\frac{d}{dt}h_j^0(t)=-A_0^Th_j^0(t)-h_j^1(t, 0)+K^{00T}(t)\phi_j(t),\\ &h_j(T)=0. \end{aligned} \tag{5.13}$$

Then using the explicit representation for $S^*(T-t)\tilde{Q}^{1/2}(e_j, 0)$ (cf. (5.24)-(5.25)) we find that

$$\psi_j^0(t) = Z_j^0(t) - \int_t^T \exp\{A_0^T(\sigma - t)\} Z_j^1(\sigma, 0)\, d\sigma + \int_t^T \exp\{A_0^T(\sigma - t)\}(\tilde{\phi}_j^T(\sigma, 0) - K^{00T}(\sigma)\phi_j(\sigma))\, d\sigma.$$

But we observe that

$$\begin{aligned}\frac{d}{dt}\left(Z_j^0(t) - \int_t^T \exp\{A(\sigma - t)\} Z_j^1(\sigma, 0)\, d\sigma\right) &= \frac{d}{dt} Y^T(T-t)Q^{1/2}e_j + Z_j^1(t, 0) + A_0^T \int_t^T \exp\{A(\sigma - t)\} Z_j^1(\sigma, 0)\, d\sigma \\ &= -A_0^T\left(Z_j^0(t) - \int_t^T \exp\{A_0^T(\sigma - t)\} Z_j^1(\sigma, 0)\, d\sigma\right).\end{aligned}$$

Consequently,

$$Z_j^0(t) - \int_t^T \exp\{A_0^T(\sigma - t)\} Z_j^1(\sigma, 0)\, d\sigma = \exp\{A_0^T(T-t)\} Q^{-1/2} e_j.$$

Hence,

$$\frac{d}{dt}\psi_j^0(t) = -A_0^T\psi_j^0(t) - \tilde{\phi}_j^T(t, 0) + K^{00T}(t)\phi_j(t), \qquad \psi_j(T) = Q^{1/2}e_j.$$

Substituting (4.8) and (5.12) into this equation with the identity $\phi_j(t) = B^T\psi_j^0(t)$ (from (5.10)) then yields

$$\begin{aligned}\frac{d}{dt}\psi_j^0(t) &= -[A_0^T - P^T(t, t)B^T]\psi_j^0(t) - \sum_{i=1}^{\nu} A_i^T B^{-T}\phi_j(t, t + r_i) \\ &\quad - \int_{-r}^0 A^T(\sigma) B^{-T}\phi_j(t, t - \sigma)\, d\sigma, \qquad (5.14) \\ \psi_j^0(T) &= Q^{1/2}e_j.\end{aligned}$$

Now we use Theorem 5.2. Multiplying the equation for $\phi(t, s)$ on the left by B^{-T} (recalling the parenthetical comment in Claim 2) and multiplying the equation for $W^T(s, t)$ on the right by B^{-1}, we obtain, upon using Proposition 5.1 and the substitution $\underset{\sim}{\phi}(t, s) = B^{-T}\phi(t, s)$,

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right) P(t, s) = -B^T\psi^0(t)\underset{\sim}{\phi}^T(t, s), \qquad P(t, T) = 0, \qquad (5.15)$$

$$\underset{\sim}{\phi}(t, s) = \psi^0(s) - \int_t^s P^T(\sigma, s) B^T\psi^0(\sigma)\, d\sigma, \qquad (5.16)$$

where $\psi^0(s)$ is the $n \times n$ matrix with jth column equal to $\psi_j^0(s)$. And with the same substitution (5.14) becomes

$$\begin{aligned}\frac{d}{dt}\psi^0(t) &= -[A_0^T - P^T(t, t)B^T]\psi^0(t) - \sum_{i=1}^{\nu} A_i^T \underset{\sim}{\phi}(t, t + r_i) \\ &\quad - \int_{-r}^0 A^T(\sigma)\underset{\sim}{\phi}(t, t - \sigma)\, d\sigma, \qquad (5.17) \\ \psi^0(T) &= Q^{1/2}.\end{aligned}$$

Next we will show that these equations are also valid when B is not invertible. Let ε_k denote a sequence of real numbers converging to zero such that $B+\varepsilon_k I$ is invertible for each k. Let B_k denote the matrix $B+\varepsilon_k I$, and consider the sequence of control problems defined by replacing B with B_k in (4.1). Introducing subscripts in the obvious fashion, we first note that since (cf. (4.5))

$$K_{(k)}(t, s) = \int_{\max(t,s)}^{T} B_k^T Y^T(\sigma - t) Q Y(\sigma - s) B_k \, d\sigma,$$

it follows that

$$K_{(k)}(t, s) - K(t, s) = \varepsilon_k \int_{\max(t,s)}^{T} Y^T(\sigma - t) Q Y(\sigma - s) B + B^T Y^T(\sigma - t) Q Y(\sigma - s) \, d\sigma.$$

Hence, it follows that

$$\sup_{t,s} |K_{(k)}(t, s) - K(t, s)| = O(\varepsilon_k), \tag{5.18}$$

and

$$\sup_{t,s} \left\{ \left| \frac{\partial}{\partial t} (K_{(k)}(t, s) - K(t, s)) \right|, \left| \frac{\partial}{\partial s} (K_{(k)}(t, s) - K(t, s)) \right| \right\} = O(\varepsilon_k).$$

Now Proposition 3.4 of [26] implies

$$\sup_{t,s} |W_{(k)}(t, s) - W(t, s)| = O(\varepsilon_k). \tag{5.19}$$

Next we consider the resolvent identities

$$W(t, s) = -X(t, s) - \int_s^t X(t, r) W(r, s) \, dr,$$

$$W(t, s) = -X(t, s) - \int_s^t W(t, r) X(r, s) \, dr$$

(and their (k) subscripted counterparts). Differentiating the top resolvent equation with respect to t and the bottom equation with respect to s, we have that Corollary 3.5 together with the estimate (5.19) lead to

$$\sup_{t,s} \left\{ \left| \frac{\partial}{\partial t} (W_{(k)}(t, s) - W(t, s)) \right|, \left| \frac{\partial}{\partial s} (W_{(k)}(t, s) - W(t, s)) \right| \right\} = O(\varepsilon_k). \tag{5.20}$$

The above estimates and the definition of $P(t, s)$ from Theorem 4.1 then imply

$$\sup_{t,s} |P_{(k)}(t, s) - P(t, s)| = O(\varepsilon_k), \tag{5.21}$$

$$\sup_{t,s} \left\{ \left| \frac{\partial}{\partial t} (P_{(k)}(t, s) - P(t, s)) \right|, \left| \frac{\partial}{\partial s} (P_{(k)}(t, s) - P(t, s)) \right| \right\} = O(\varepsilon_k). \tag{5.22}$$

Now consider the (k)-subscripted equations in (5.15)–(5.17) (which are valid since B_k is invertible):

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial s} \right) P_{(k)}(t, s) = -B_k^T \psi_{(k)}^0(t) \underset{\sim}{\phi}_{(k)}^T(t, s), \qquad P_{(k)}(t, T) = 0, \tag{5.15'}$$

$$\underset{\sim}{\phi}_{(k)}(t,s)=\psi^0_{(k)}(s)-\int_t^s P^T_{(k)}(\sigma,s)B^T_k\psi^0_{(k)}(\sigma)\,d\sigma, \tag{5.16'}$$

$$\frac{d}{dt}\psi^0_{(k)}(t)=-[A_0^T-P^T_{(k)}(t,t)B_k^T]\psi^0_{(k)}(t)-\sum_{i=1}^{\nu}A_i^T\underset{\sim}{\phi}_{(k)}(t,t+r_i)$$

$$-\int_{-r}^0 A^T(\sigma)\underset{\sim}{\phi}_{(k)}(t,t-\sigma)\,d\sigma,$$

$$\psi^0_{(k)}(T)=Q^{1/2}.$$

From (5.5) and (5.19) it follows that $\phi_{(k)}(t)\to\phi(t)$ uniformly on $[0,T]$. And since $P_{(k)}(t,s)\to P(t,s)$ uniformly, (5.8)-(5.9) and (4.8)-(4.9) imply that $\psi_{(k)}(t)\to\psi(t)$ uniformly; and in particular $\psi_{(k)}(t)\to\psi^0(t)$ in $R^{n\times n}$ on $[0,T]$. Thus

$$\lim_{k\to\infty}\underset{\sim}{\phi}_{(k)}(t,s)=\psi^0(s)-\int_t^s P^T(\sigma,s)B^T\psi^0(\sigma)\,d\sigma. \tag{5.23}$$

Denote this limit by $\underset{\sim}{\phi}(t,s)$. Next take limits on both sides of (5.23) and (5.24), and use the convergence properties of $\psi^0_{(k)}$, $\underset{\sim}{\phi}_{(k)}$, and $(\partial/\partial t+\partial/\partial s)P_{(k)}(t,s)$ (from (5.22)) to obtain

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial s}\right)P(t,s)=-B^T\psi^0(t)\underset{\sim}{\phi}^T(t,s),\qquad P(t,T)=0,$$

$$\frac{d}{dt}\psi^0(t)=-[A_0^T-P^T(t,t)B^T]\psi^0(t)-\sum_{i=1}^{\nu}A_i^T\underset{\sim}{\phi}(t,t+r_i)-\int_{-r}^0 A^T(\sigma)\underset{\sim}{\phi}(t,t-\sigma)\,d\sigma.$$

Taking derivatives in (5.23), and removing the undertilde on $\underset{\sim}{\phi}(\cdot,\cdot)$ results in

$$\frac{\partial}{\partial t}\phi(t,s)=P^T(t,s)B^T\psi^0(t),\qquad \phi(s,s)=\psi^0(s),$$

thereby establishing the equations of the theorem.

The differentiability properties of P are stated in Theorem 4.1, those of ψ^0 follow from (5.17'), (5.16'), and (5.23), and the properties of ϕ are obtained from (5.23).

To prove uniqueness of the solution, it suffices to demonstrate it locally, since global existence has already been established. This matter is facilitated by the use of the equivalent integral equations (6.4)-(6.7) in the following section. This formulation shows with standard arguments that $P(\cdot,\cdot)$, and $\phi(\cdot,\cdot)$ are locally propagated via a contraction map (cf. (6.26)-(6.31)); hence we obtain local (and consequently, global) uniqueness.

Proof of Claim 1. Using the representation (4.8)-(4.9), we have

$$K^*(t)\phi_j(t)=\Bigg(P^T(t,t)\phi_j(t),\ \sum_{i=1}^{\nu}\chi[-r_i,0](\cdot)A_i^TP(t,\cdot+t+r_i)\phi_j(t)$$

$$+\int_t^{t+r}A^T(\cdot+t-\sigma)P^T(t,\sigma)\,d\sigma\,\phi_j(t)\Bigg).$$

From here it is straightforward to verify that $K^*(t)\phi^*(t)\in D(A^*)$ for $t<T$. Using the differentiability properties of $P(t,\alpha)$ and the definition of A^*, we then compute

$$A^*K^*(t)\phi_j(t)=\Bigg(A_0^TP^T(t,t)\phi_j(t)+\sum_{i=1}^{\nu}A_i^TP^T(t,t+r_i)\phi_j(t)$$

$$+\int_t^{t+r}A^T(t-\sigma)P^T(t,\sigma)\phi_j(t)\,d\sigma,\ A^T(\theta)P^T(t,t)\phi_j(t)$$

$$-\sum_{i=1}^{\nu}\chi[-r_i,0](\theta)A_i^T\frac{\partial}{\partial\theta}P(t,\theta+t+r_i)\phi_j(t)$$

$$-\frac{\partial}{\partial\theta}\left\{\int_{-r}^{\theta}A^T(s)P^T(t,\theta+t-s)\,ds\right\}\Bigg).$$

Hence, $t\to A^*(t)K^*(t)\phi_j(t)$ is seen to be an $L_1(0,T;M_2)$-valued function. The proof of the claim then follows from [33, p. 108].

Proof of Claim 2. We will first need an explicit characterization of $S^*(T-t)\tilde{Q}^{1/2}(e_j,0)$. Define the F-structural operator [6] via the relation

$$F(\alpha,x(\cdot))=\left(\alpha,\sum_{i=1}^{\nu}A_i^Tx(-r_i-\cdot)\chi[-r_i,0](\cdot)+\int_{-r}^{\cdot}A^Tx(s-\cdot)\,ds\right)$$

for $(\alpha,x(\cdot))\in M_2$. The following identity is easily shown to hold [6]:

$$S^*(T-t)\tilde{Q}^{1/2}=S^*(T-t)F\tilde{Q}^{1/2}=FS_T(T-t)\tilde{Q}^{1/2},$$

where

$$S_T(T-t)\tilde{Q}^{1/2}(e_j,0)=(Y^T(T-t)Q^{1/2}e_j,\ Y^T(T-t+\cdot)Q^{1/2}e_j).$$

Now write $S^*(T-t)\tilde{Q}^{1/2}(e_j,0)$ in component form as $(Z_j^0(t),Z_j^1(t,\cdot))$. Using the relations above we then obtain

$$Z_j^0(t)=Y^T(T-t)Q^{1/2}e_j, \tag{5.24}$$

$$Z_j^1(t,\alpha)=\left\{\sum_{i=1}^{\nu}A_i^TY^T(T-t-r_i-\alpha)\chi[-r_i,0](\alpha)\right.$$

$$\left.+\int_{-r}^{\alpha}A^T(s)Y^T(T-t+s-\alpha)\,ds\right\}Q^{1/2}e_j. \tag{5.25}$$

Now note that

$$\psi_j^0(T)=Q^{1/2}e_j,\qquad \tilde{\phi}_j(T,\alpha)=0, \tag{5.26}$$

for $t<T$,

$$\tilde{\phi}_j(t,-r_\nu)=\phi_j^T(t)B^{-1}A_\nu,$$

$$\tilde{\phi}_j(t,-r_i)^+-\tilde{\phi}_j(t,-r_i)^-=\tilde{\phi}_j^TB^{-1}A_i,\qquad i=1,\cdots,\nu-1.$$

Since $\phi_j(t)=B^T\psi_j^0(t)$ (from (5.10)),

$$\tilde{\phi}_j^T(t,-r_\nu)=A_\nu^T\psi_j^0(t), \tag{5.27}$$

$$\tilde{\phi}_j^T(t,-r_i)^+-\tilde{\phi}(t,-r_i)^-=A_i^T\psi_j^0(t),\qquad i=1,\cdots,\nu-1. \tag{5.28}$$

From (5.24)-(5.25) we also compute

$$Z_j^0(T)=Q^{1/2}e_j,\qquad Z_j^1(T,\alpha)=0, \tag{5.29}$$

$$Z_j^1(t,-r_\nu)=A_\nu^TY^T(T-t)Q^{1/2}e_j=A_\nu^TZ_j^0(t), \tag{5.30}$$

$$Z_j^1(t,-r_i)^+-Z_j^1(t,-r_i)^-=A_i^TY^T(T-t)Q^{1/2}e_j=A_j^TZ_j^0(t), \tag{5.31}$$

$$i=1,\cdots,\nu-1.$$

Next define $d(t)\in M_2$ to be the right side of (5.12), and let $d^0(t)$ and $d^1(t)$ denote the R^n and $L_2(-r,0;R^n)$ components of $d(t)$, respectively. Then (5.26)-(5.31) imply

$$d^0(T)=d^1(T,\alpha)=0, \tag{5.32}$$

$$d^1(t,-r_\nu)=A_\nu^Td^0(t),\qquad t\in[0,T], \tag{5.33}$$

$$d^1(t,-r_i)^+-d^1(t,-r_i)^-=A_i^Td^0(t),\qquad t\in[0,T),\quad i=1,\cdots,\nu-1. \tag{5.34}$$

From the definition of $\phi_j(\cdot)$ and the fact that $W(\cdot,\cdot)$ has continuous first partial derivatives, it follows that for fixed t, $t \leqq T-r$, $\tilde{\phi}_j(t,\alpha)$ has bounded α-derivatives in each interval $(-r_{i+1}, -r_i)$. And since $Z_j^1(t,\alpha)$ also has bounded α-derivatives for $t \leqq T-r$ on each interval $(-r_{i+1}, -r_i)$, (5.33)-(5.34) imply that $d(t) \in D(A^*)$ for $t \leqq T-r$. Again referring to (5.33)-(5.34), it will be established that $d(t) \in D(A^*)$ for $t \leqq T$ once it is shown that for each $t \in (T-r, T)$, $d^1(t,\cdot) \in H^1(-r_{i+1}, -r_i)$.

So fix $t_0 \in (T-r, T)$ and consider $Z_j^1(t_0,\alpha)$. For $\alpha \in (-r_{i+1}, -r_i)$, $Z_j^1(t_0,\alpha)$ has a bounded α-derivative except possibly at points $\alpha_k = T - t_0 - r_k$, $k = i+1, \cdots, \nu$. On the other hand, looking at the α-derivative of $\tilde{\phi}_j(t_0,\alpha)$ for $\alpha \in (-r_{i+1}, -r_i)$, we also find a bounded derivative, except possibly at $\{\alpha_k\}_{k=i+1}^{\nu}$. Thus it suffices to show that $d^1(t,\cdot)$ is continuous in $(-r_{i+1}, -r_i)$. Since the only possible discontinuities in $d^1(t_0,\cdot)$ occur at the α_k's, we compute $d^1(t_0,\alpha_k)^+ - d^1(t_0,\alpha_k)^-$:

$$\begin{aligned} d^1(t_0,\alpha_k)^+ - d^1(t_0,\alpha_k)^- &= \tilde{\phi}_j^T(t_0,\alpha_k)^+ - \tilde{\phi}_j^T(t_0,\alpha_k)^- - (Z_j^1(t_0,\alpha_k)^+ - Z_j^1(t_0,\alpha_k)^-) \\ &= A_k^T B^{-T}\phi_j(T) - A_k^T Q^{1/2} \\ &= 0 \end{aligned}$$

by (5.7). Therefore we conclude that $d(t) \in D(A^*)$, $t \leqq T$.

Next we apply the differential operator $D = \partial/\partial t - \partial/\partial\alpha$ to $d^1(t,\alpha)$. For $\alpha \in (-r_{i+1}, -r_i)$, straightforward computations using (5.12) and (5.24)-(5.25) yield

$$\begin{aligned} D(d^1(t,\alpha)) = &-\sum_{k=i+1}^{\nu} A_k^T B^{-T} W(\alpha + r_k + t, t)\phi_j(t) \\ &- A^T(\alpha)B^{-T}\phi_j(t) + A^T(\alpha)Y^T(T-t)Q^{1/2}e_j \\ &- \int_{-r}^{\alpha} A^T(\sigma)B^{-T}W(\alpha - \sigma + t, t)\phi_j(t)\, d\sigma \\ = &-A^T(\alpha)[B^{-T}\phi_j(t) - Y^T(T-t)Q^{1/2}e_j] \\ &- \left\{\int_{-r}^{\alpha} A^T(\sigma)B^{-T}W(\alpha - \sigma + t, t)\right\} d\sigma \\ &+ \sum_{k=i+1}^{\nu} A_k^T B^{-T} W(\alpha + r_k + t, t)\phi_j(t). \end{aligned}$$

Now (5.10) implies $B^{-T}\phi_j(t) = \psi_j^0(t)$, so that $B^{-T}\phi_j(t) - Y^T(T-t)Q^{1/2}e_j = d^0(t)$. Thus using (4.8)-(4.9) and Proposition 5.1, we arrive at

$$D(d^1(t,\alpha)) = -A^T(\alpha)d^0(t) + K^{01^T}(t,\alpha)\phi_j(t). \tag{5.35}$$

Definition (5.9) implies $d^0(t) = h_j^0(t)$ (where we have written $h_j(t) = (h_j^0(t), h_j^1(t))$). Let $r(t) = d(t) - h_j(t)$. Then $r(t) = (r^0(t), r^1(t,\cdot)) \in D(A^*)$ with $r^0(t) = d^0(t) - h_j^0(t) = 0$. Furthermore, from (5.13), (5.32)-(5.35) it follows that

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial\alpha}\right) r^1(t,\alpha) = 0$$

with boundary condition

$$r^1(T,\alpha) = 0.$$

Hence $r^1(t,\alpha) = 0$, so that $d^1(t,\alpha) = h_j^1(t,\alpha)$. And indeed $d(t) = h_j(t)$. This completes the proof of Claim 2 and the theorem. □

Observe that in the case of no delay terms the kernel $P(t, \alpha)$ needs only to be specified on the diagonal, i.e.,

$$u(t) = -P(t, t)x(t).$$

The relevant equations (5.3)-(5.4) then collapse to

$$\frac{d}{dt}P(t, t) = -B^T\psi^0(t)\psi^{0T}(t), \qquad P(T, T) = 0,$$

$$\frac{d}{dt}\psi^0(t) = -[A_0^T - P^T(t, t)B^T]\psi^0(t), \qquad \psi^0(T) = Q^{1/2}.$$

These equations are recognized as the Chandrasekhar equations for systems without delay. Thus the full set of equations (5.3)-(5.4) can be interpreted as an explicit generalization of the finite-dimensional Chandrasekhar equations to systems with delay.

6. The structure of the gain. In this section we analyze the structure of the optimal gain subject to appropriate regularity conditions. We consider the system (5.3)-(5.4) and introduce the following changes of variables:

$$\begin{aligned} &(t, s) \to (t, t+\mu), \quad \tilde{\phi}(t, \mu) = \phi(t, t+\mu), \quad \tilde{P}(t, \mu) = P(t, t+\mu), \\ &\tilde{\phi}_i(t) = \tilde{\phi}(t, r_i), \quad i = 0, \cdots, \nu, \qquad \tilde{P}_i(t) = \tilde{P}(t, r_i), \quad i = 0, \cdots, \nu. \end{aligned} \tag{6.1}$$

We note that with this notation $\tilde{\phi}_0 = \psi^0$, where ψ^0 is defined in Theorem 5.3. The transformation of (5.3)-(5.4) becomes

$$\frac{\partial}{\partial t}\tilde{P}(t, \mu) = -B^T\tilde{\phi}_0(t)\tilde{\phi}^T(t, \mu), \tag{6.2a}$$

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial \mu}\right)\tilde{\phi}(t, \mu) = \tilde{P}^T(t, \mu)B^T\tilde{\phi}_0(t), \tag{6.2b}$$

$$\frac{d}{dt}\tilde{\phi}_0(t) = -[A_0^T - \tilde{P}_0^T(t)B^T]\tilde{\phi}_0(t) - \sum_{i=1}^{\nu} A_i^T\tilde{\phi}_i(t) - \int_0^r A^T(\theta)\tilde{\phi}(t, \theta)\, d\theta, \tag{6.2c}$$

with the initial conditions and history conditions:

$$\begin{aligned} &\tilde{P}(t, \mu) = 0, && t+\mu = T, \\ &\tilde{\phi}(t, \mu) = Q^{1/2}, && t+\mu = T, \\ &\tilde{\phi}(t, \mu) = 0, && t+\mu > T. \end{aligned} \tag{6.3}$$

Now we reformulate (6.2)-(6.3) as a self-contained system of integral equations. The equation for $\tilde{\phi}(t, \mu)$ is hyperbolic and can be integrated by the method of characteristics. The equations for $\tilde{P}(t, \mu)$ and $\tilde{\phi}_0(t)$ can be explicitly integrated from the initial conditions given by (6.3). Thus, we have

$$\tilde{P}(t, \mu) = -\int_{T-\mu}^{t} B^T\tilde{\phi}_0(\tau)\tilde{\phi}^T(\tau, \mu)\, d\tau, \tag{6.4a}$$

$$\begin{aligned} \tilde{\phi}_0(t) = Q^{1/2} - \int_T^t \Bigg\{ &[A_0^T - \tilde{P}_0^T(\tau)B^T]\tilde{\phi}_0(\tau) + \sum_{i=1}^{\nu} A_i^T\tilde{\phi}_i(\tau) \\ &+ \int_0^r A^T(\theta)\tilde{\phi}(\tau, \theta)\, d\theta \Bigg\} d\tau, \end{aligned} \tag{6.4b}$$

where we have defined

$$\tilde{\phi}(t, \mu) = \tilde{\phi}_0(t+\mu) - \int_t^{t+\mu} \tilde{P}^T(\sigma, t+\mu-\sigma)B^T\tilde{\phi}_0(\sigma)\, d\sigma, \tag{6.5a}$$

$$\tilde{\phi}_i(t) = \tilde{\phi}(t, r_i) = \lim_{\mu \to r_i} \tilde{\phi}(t, \mu), \tag{6.5b}$$

$$\tilde{P}_i(t) = \tilde{P}(t, r_i) = \lim_{\mu \to r_i} \tilde{P}(t, \mu). \tag{6.5c}$$

The functions $\tilde{P}(t, \mu)$ and $\tilde{\phi}_0(t)$ are, respectively, defined on the sets:

$$\begin{aligned} &D = \{(t, \mu) \mid t > 0,\ t + \mu < T,\ 0 < \mu < r\}, \\ &D_0 = \{t \mid 0 < t < T\}. \end{aligned} \tag{6.6}$$

This formulation is completed by the history condition

$$\begin{aligned} &\tilde{\phi}(t, \mu) = 0, \qquad (t, \mu) \in D', \\ &D' = \{(t, \mu) \mid t + \mu > T,\ t < T,\ 0 < \mu < r\}. \end{aligned} \tag{6.7}$$

It will also be convenient to represent the delay interval as

$$D_r = \{\mu \mid 0 < \mu < r\}. \tag{6.8}$$

The integral system (6.4)–(6.7) is very useful for clarifying the smoothness of the gain. To this end, we introduce the following notation. Let N_1 and N_2 be given positive integers and let Ω denote a subset of R^{N_1}. The interior of Ω is denoted by $\mathring{\Omega}$ and the closure of Ω is denoted by $\bar{\Omega}$. Let $C(\Omega)$ denote the space of continuous functions from Ω into R^{N_2}. If Ω is open and k is a positive integer, let $C^k(\Omega)$ denote the space of continuous functions possessing continuous derivatives up to order k on Ω, and let $C^k(\bar{\Omega})$ denote the space of all $u \in C(\Omega)$ such that all derivatives of order k or less extend continuously to $\bar{\Omega}$. And also we define

$$C^\infty(\Omega) = \bigcap_{k=1}^{\infty} C^k(\Omega), \qquad C^\infty(\bar{\Omega}) = \bigcap_{k=1}^{\infty} C^k(\bar{\Omega}).$$

The appropriate values of N_1 and N_2 will always be clear from the context in which this notation is used. In addition we introduce special notation for one and two dimensions. Let $f(t)$ and $g(t, \mu)$ be functions of t and (t, μ), respectively, on prescribed domains D_f and D_g in R^1 and R^2, respectively. Differentiation is denoted by

$$f(t)^{[n]} = \frac{d^n}{dt^n} f(t),$$

$$g(t, \mu)^{[n_1, n_2]} = \left(\frac{\partial^{n_1 + n_2}}{\partial t^{n_1} \partial \mu^{n_2}} \right) g(t, \mu).$$

We now suppose that D_f can be subdivided into a finite number of open sets D_f^i where

$$D_f^i \subset D_f, \qquad D_f \subset \left(\bigcup_{i=1}^{N_f} \bar{D}_f^i \right).$$

We suppose further that $f(t)$ is in $C^p(\bar{D}_f^i)$ for each $i \in \{1, \cdots, N_f\}$ and that, for $p > 0$, $f(t)$ is in $C^{p-1}(\bar{D}_f)$. Then $f(t)$ is said to be piecewise $C^p[t]$ in D_f. Let $\hat{K}$ be a finite set of points that determine such a subdivision. Then $f(t)$ is said to be piecewise $C^p[t]$ in D_f with respect to $\hat{K}$.

Similarly, we first assume that D_g can be subdivided into a finite number of open sets D_g^i where

$$D_g^i \subset D_g, \qquad D_g \subset \left(\bigcup_{i=1}^{N_g} \bar{D}_g^i \right).$$

We suppose that $g(t, \mu)$ is in $C^p(\bar{D}_g^i)$ for each $i \in \{1, \cdots, N_g\}$ and that, for $p > 0$, $g(t, \mu)$ is in $C^{p-1}(\bar{D}_g)$. Then $g(t, \mu)$ is said to be piecewise $C^p[t, \mu]$ in D_g. And also let $\hat{L}$ be a finite set of line segments that determine such a subdivision. Then $g(t, \mu)$ is said to be piecewise $C^p[t, \mu]$ in D_g with respect to $\hat{L}$.

The following theorem, which is illustrated in Fig. 1 for the case $\nu = 3$, outlines the structure of the optimal gain and leads to the development of second-order difference approximations in § 7. It is convenient to define the following subdomains:

$$\hat{D}^i = \{(t, \mu) \mid T - r_{i+1} < t + \mu < T - r_i\} \cap D, \qquad i \in \{0, \cdots, \nu - 1\},$$

$$\hat{D}^\nu = \{(t, \mu) \mid 0 < t + \mu < T - r_\nu\} \cap D.$$

THEOREM 6.1. *Let the solution to the system* (6.4)-(6.7) *be given by* $\tilde{P}(t, \mu)$ *and* $\tilde{\phi}_0(t)$. *The smoothness of the solution is characterized as follows*:

(1) $\tilde{P}(t, \mu)$ *is piecewise* $C^2[t, \mu]$ *in* D *with respect to* $\{(t, \mu) \mid t = T - r_i\}_{i=1}^{\nu}$, $\{(t, \mu) \mid \mu = r_i\}_{i=1}^{\nu}$, *and* $\{(t, \mu) \mid t + \mu = T - r_i\}_{i=1}^{\nu}$.

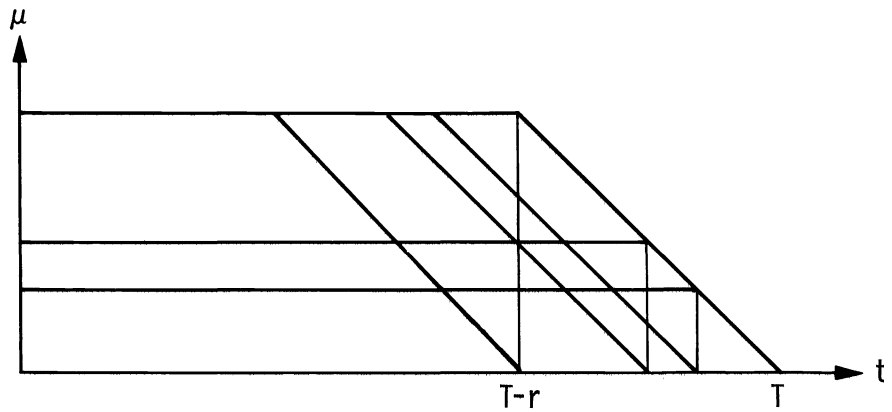


FIG. 1. *Structure for multiple delays.*

(2) $\tilde{\phi}(t, \mu)$ *is piecewise* $C^1[t, \mu]$ *in* D *with respect to* $\{(t, \mu) \mid t + \mu = T - r_i\}_{i=1}^{\nu}$. *And, moreover, for each* $i \in \{0, \cdots, \nu\}$, $\tilde{\phi}(t, \mu)$ *is piecewise* $C^2[t, \mu]$ *in* $\hat{D}^i$ *with respect to* $\{(t, \mu) \mid t + \mu = T - r_j - r_k\}_{j=1,k=0}^{\nu,\nu}$.

Proof. By Theorem 5.3 we have that $\tilde{\phi}_0(t)$ is piecewise $C^1[t]$ in D_0 and that $\tilde{P}(t, \mu)$ is in $C^1(\bar{D})$. Then by the definition (6.5a) we have that $\tilde{\phi}(t, \mu)$ is piecewise $C^1[t, \mu]$ in D. Using the integral formulation (6.4)-(6.7), we formally generate the relevant first-order derivatives. Our formulas are recursive in that a given expression for a derivative may depend on previously given expressions. We have

$$\tilde{\phi}_0(t)^{[1]} = -\left\{ [A_0^T - \tilde{P}_0^T(t) B^T] \tilde{\phi}_0(t) + \sum_{i=1}^{\nu} A_i \tilde{\phi}_i(t) + \int_0^r A^T(\theta) \tilde{\phi}(t, \theta)\, d\theta \right\}, \tag{6.9a}$$

$$\tilde{P}(t, \mu)^{[1,0]} = -B^T \tilde{\phi}_0(t) \tilde{\phi}^T(t, \mu), \tag{6.9b}$$

$$\tilde{P}(t,\mu)^{[0,1]} = -B^T\tilde{\phi}_0(T-\mu)\tilde{\phi}^T(T-\mu,\mu) - \int_{T-\mu}^{t} B^T\tilde{\phi}_0(\tau)\tilde{\phi}^T(\tau,\mu)^{[0,1]}\,d\tau$$

$$= -B^T\tilde{\phi}_0(T-\mu)\tilde{\phi}^T(T-\mu,\mu)$$

$$- \int_{T-\mu}^{t} B^T\tilde{\phi}_0(\tau)[\tilde{\phi}^T(\tau,\mu)^{[1,0]} - \tilde{\phi}_0^T(\tau)B\tilde{P}(\tau,\mu)]\,d\tau \tag{6.9c}$$

$$= -B^T\tilde{\phi}_0(t)\tilde{\phi}^T(t,\mu) + \int_{T-\mu}^{t} B^T\tilde{\phi}_0(\tau)^{[1]}\tilde{\phi}^T(\tau,\mu)\,d\tau$$

$$+ \int_{T-\mu}^{t} B^T\tilde{\phi}_0(\tau)\tilde{\phi}_0^T(\tau)B\tilde{P}(\tau,\mu)\,d\tau.$$

We note that (6.4b) was used to simplify expression (6.9c) by eliminating the μ-derivative of $\tilde{\phi}(t,\mu)$. The subsequent integration by parts with respect to t is justified since $\tilde{\phi}(t,\mu)$ is piecewise $C^1[t,\mu]$ in D. We now complete the enumeration of first-order derivatives:

$$\tilde{\phi}(t,\mu)^{[0,1]} = \tilde{\phi}_0(t+\mu)^{[1]} - \tilde{P}_0^T(t+\mu)B^T\tilde{\phi}_0(t+\mu)$$

$$- \int_{t}^{t+\mu} \tilde{P}^T(\sigma, t+\mu-\sigma)^{[0,1]}B^T\tilde{\phi}_0(\sigma)\,d\sigma, \tag{6.9d}$$

$$\tilde{\phi}(t,\mu)^{[1,0]} = \tilde{\phi}(t,\mu)^{[0,1]} + \tilde{P}^T(t,\mu)B^T\tilde{\phi}_0(t). \tag{6.9e}$$

And finally, from (6.9d) and (6.9e) we have for $i \in \{1, \cdots, \nu\}$:

$$\tilde{\phi}_i^{[1]}(t) = \tilde{\phi}_0(t+r_i)^{[1]} - \tilde{P}_0^T(t+r_i)B^T\tilde{\phi}_0(t+r_i) + \tilde{P}_i^T(t)B^T\tilde{\phi}_0(t)$$

$$- \int_{t}^{t+r_i} \tilde{P}^T(\sigma, t+r_i-\sigma)^{[0,1]}B^T\tilde{\phi}_0(\sigma)\,d\sigma. \tag{6.9f}$$

Similarly, the second-order derivatives are formally given by

$$\tilde{\phi}_0(t)^{[2]} = -\Big\{[A_0^T - \tilde{P}_0^T(t)B^T]\tilde{\phi}_0(t)^{[1]} - \tilde{P}_0(t)^{[1]}B^T\tilde{\phi}_0(t)$$

$$+ \sum_{i=1}^{\nu} A_i\tilde{\phi}_i(t)^{[1]} + \int_0^r A^T(\theta)\tilde{\phi}(t,\theta)^{[1,0]}\,d\theta\Big\}, \tag{6.10a}$$

$$\tilde{P}(t,\mu)^{[2,0]} = -B^T\tilde{\phi}_0(t)^{[1]}\tilde{\phi}^T(t,\mu) - B^T\tilde{\phi}_0(t)\tilde{\phi}^T(t,\mu)^{[1,0]}, \tag{6.10b}$$

$$\tilde{P}(t,\mu)^{[1,1]} = -B^T\tilde{\phi}_0(t)\tilde{\phi}^T(t,\mu)^{[0,1]}, \tag{6.10c}$$

$$\tilde{P}(t,\mu)^{[0,2]} = -B^T\tilde{\phi}_0(t)\tilde{\phi}(t,\mu)^{[0,1]} + B^T\tilde{\phi}_0(T-\mu)^{[1]}Q^{1/2}$$

$$+ \int_{T-\mu}^{t} B^T\tilde{\phi}_0(\tau)^{[1]}\tilde{\phi}^T(\tau,\mu)^{[0,1]}\,d\tau \tag{6.10d}$$

$$+ \int_{T-\mu}^{t} B^T\tilde{\phi}_0(\tau)\tilde{\phi}_0^T(\tau)B\tilde{P}(\tau,\mu)^{[0,1]}\,d\tau,$$

$$\tilde{\phi}(t,\mu)^{[0,2]} = \tilde{\phi}(t+\mu)^{[2]} - \tilde{P}_0^T(t+\mu)B^T\tilde{\phi}_0(t+\mu)^{[1]} - \tilde{P}_0^T(t+\mu)^{[1]}B^T\tilde{\phi}_0(t+\mu)$$

$$-\tilde{P}^T(t+\mu,0)^{[0,1]}B^T\tilde{\phi}_0(t+\mu)$$

$$- \int_{t}^{t+\mu} \tilde{P}^T(\sigma, t+\mu-\sigma)^{[0,2]}B^T\tilde{\phi}_0(\sigma)\,d\sigma, \tag{6.10e}$$

$$\tilde{\phi}(t,\mu)^{[1,1]} = \tilde{\phi}(t,\mu)^{[0,2]} + \tilde{P}^T(t,\mu)^{[0,1]}B^T\tilde{\phi}_0(t), \tag{6.10f}$$

$$\tilde{\phi}(t,\mu)^{[2,0]} = \tilde{\phi}(t,\mu)^{[1,1]} + \tilde{P}^T(t,\mu)^{[1,0]}B^T\tilde{\phi}_0(t) + \tilde{P}^T(t,\mu)B^T\tilde{\phi}_0(t)^{[1]}. \tag{6.10g}$$

By means of recursively evaluating the formulas (6.9)-(6.10) we now proceed sequentially to outline the differentiability of $\tilde{P}(t,\mu)$ and $\tilde{\phi}(t,\mu)$. We begin by noting that $\tilde{\phi}_0(t)$ and $\tilde{P}(t,\mu)$ are bounded and continuous in D_0 and D, respectively. On the other hand, for $i\in\{1,\cdots,\nu\}$, $\tilde{\phi}_i(t)$ is discontinuous in D_0 at $t=T-r_i$. From this we can derive global results. We have:

(1) $\tilde{\phi}_0^{[1]}$ is piecewise $C^0[t]$ in D_0 with respect to $\{T-r_i\}_{i=1}^{\nu}$;

(2) $\tilde{P}^{[1,0]}$ and $\tilde{P}^{[0,1]}$ are in $C^0(\bar{D}_0)$;

(3) $\tilde{\phi}^{[0,1]}$ and $\tilde{\phi}^{[1,0]}$ are piecewise $C^0[t,\mu]$ in D with respect to $\{(t,\mu)\,|\,t+\mu=T-r_i\}_{i=1}^{\nu}$;

(4) For $i\in\{1,\cdots,\nu\}$, $\tilde{\phi}_i^{[1]}$ is piecewise $C^0[t]$ in $\{t\,|\,0<t<T-r_i\}$ with respect to $\{T-r_i-r_j\}_{j=1}^{\nu}$;

(5) For each $i\in\{0,\cdots,\nu-1\}$, $\tilde{\phi}_0^{[2]}$ is piecewise $C^0[t]$ in $\{t\,|\,T-r_{i+1}<t<T-r_i\}$ with respect to $\{T-r_j-r_k\}_{j=1,k=0}^{\nu,\nu}$. And also $\phi_0^{[2]}$ is piecewise $C^0[t]$ in $\{t\,|\,0<t<T-r\}$ with respect to $\{T-r_j-r_k\}_{j=1,k=0}^{\nu,\nu}$.

(6) $\tilde{P}^{[2,0]}$ is piecewise $C^0[t,\mu]$ in D with respect to $\{(t,\mu)\,|\,t=T-r_i\}_{i=1}^{\nu}$ and $\{(t,\mu)\,|\,t+\mu=T-r_i\}_{i=1}^{\nu}$;

(7) $\tilde{P}^{[1,1]}$ is piecewise $C^0[t,\mu]$ in D with respect to $\{(t+\mu)\,|\,t+\mu=T-r_i\}_{i=1}^{\nu}$;

(8) $\tilde{P}^{[0,2]}$ is piecewise $C^0[t,\mu]$ in D with respect to $\{(t,\mu)\,|\,t+\mu=T-r_i\}_{i=1}^{\nu}$ and $\{(t,\mu)\,|\,\mu=r_i\}_{i=1}^{\nu}$.

(9) For each $i\in\{0,\cdots,\nu\}$, $\tilde{\phi}^{[0,2]}$, $\tilde{\phi}^{[1,1]}$ and $\tilde{\phi}^{[2,0]}$ are piecewise $C^0[t,\mu]$ in $\hat{D}^i$ with respect to $\{(t,\mu)\,|\,t+\mu=T-r_j-r_k\}_{j=1,k=0}^{\nu,\nu}$.

This enumeration completes the justification of the theorem. □

The preceding theorem strengthens the differentiability properties of $P(t,\mu)$ as derived in Theorem 4.1. We now demonstrate that the assumption of a single delay ($\nu=1$) and a smooth integral term leads to a considerably more detailed structure for the optimal gain. Without loss of generality we can assume that $T=Kr$, where K is a positive integer greater than one, since the interval D_0 can be translated to lie within some interval $\{t\,|\,0<t<Kr\}$ with the right endpoints aligned.

For $k\in\{0,1,2,\cdots,K\}$ we define the points

$$t_k=T-kr,\quad p_k^1=(t_k,0),\quad p_k^2=(t_k,r). \tag{6.11}$$

Then for $k\in\{0,1,2,\cdots,K-1\}$ we define the open line segments

$$\begin{aligned} L_k^0&=\{(t,\mu)\,|\,t_{k+1}<t<t_k,\ \mu=0\},\\ L_k^1&=\{(t,\mu)\,|\,t+\mu=t_k,\ 0<\mu<r\},\\ L_k^2&=\{(t,\mu)\,|\,t=t_k,\ 0<\mu<r\}, \end{aligned} \tag{6.12}$$

and the open triangles

$$\begin{aligned} T_k^1&=\{(t,\mu)\,|\,t+\mu<t_k,\ 0<\mu<r,\ t_{k+1}<t<t_k\},\\ T_k^2&=\{(t,\mu)\,|\,t+\mu>t_k,\ 0<\mu<r,\ t_{k+1}<t<t_k\}. \end{aligned} \tag{6.13}$$

These sets are illustrated in Fig. 2 for the case $T=4r$. For $k\in\{0,1,2,\cdots,K-1\}$ we also define the functions

$$\hat{\mu}_k(t)=t_k-t,\qquad \hat{t}_k(\mu)=t_k-\mu, \tag{6.14}$$

which for appropriate t and μ define reference points along the line L_k^1. The following theorem, which elucidates the structure of the gain for this important special case, leads to the development of difference methods of arbitrarily high order in § 7.

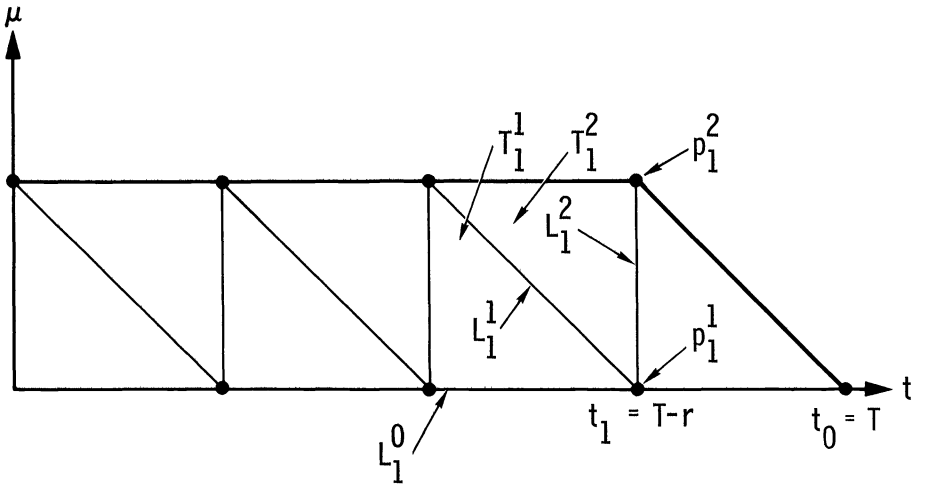


FIG. 2. *Structure for a single delay.*

THEOREM 6.2. *Let K be some positive integer greater than one. Let the solution to the system* (6.4)-(6.7) *with $\nu = 1$, $A(\Theta) \in C^\infty(\bar{D}_r)$, and $T = Kr$ be given by $\tilde{P}(t, \mu)$ and $\tilde{\phi}_0(t)$. For $k \in \{1, 2, \cdots, K-1\}$ the smoothness of the solution is characterized as follows:*

(1) *$\tilde{P}(t, \mu)$ and $\tilde{\phi}(t, \mu)$ are in $C^\infty(\bar{T}_0^1)$;*

(2) *$\tilde{P}(t, \mu)$ and $\tilde{\phi}(t, \mu)$ are in $C^\infty(\bar{T}_k^1)$ and $C^\infty(\bar{T}_k^2)$;*

(3) *Across the point p_k^2 for $k \geqq 2$, $\tilde{P}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k-1$ and $\tilde{\phi}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k-2$;*

(4) *Across the line L_k^2, $\tilde{P}(t, \mu)^{[j_1, j_2]}$ and $\tilde{\phi}(t, \mu)^{[j_1, j_2]}$ are continuous for $j_1 + j_2 \leqq k$;*

(5) *Across the point p_k^1, $\tilde{P}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k$ and $\tilde{\phi}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k-1$;*

(6) *Across the line L_k^1, $\tilde{P}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k$ and $\tilde{\phi}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k-1$.*

Proof. First we consider the solution in T_0^1. The delay has no effect in this region and (6.4b) becomes

$$\tilde{\phi}_0(t) = Q^{1/2} - \int_T^t \left\{ [A_0^T - \tilde{P}_0^T(t) B^T] \tilde{\phi}_0(\tau) + \int_0^{T-r} A(\Theta) \tilde{\phi}(\tau, \Theta)\, d\Theta \right\} d\tau.$$

Then the system of integral equations (6.4)-(6.7) can be differentiated repeatedly to guarantee that $\tilde{P}(t, \mu)$ and $\tilde{\phi}(t, \mu)$ are in $C^\infty(\bar{T}_0^1)$. By our assumptions, the solution is at least continuous across the line L_1^2, and in particular, $\tilde{\phi}(t)$ is continuous across t_1. But then $\tilde{P}_0(t)$ must have a continuous derivative across t_1 since by (6.4a) it is an integral of $\tilde{\phi}_0(t)$. On L_1^2 the solution $(\tilde{\phi}(t, \mu), \tilde{P}(t, \mu))$ must have continuous first-order derivatives since (6.4a) and (6.5a) can be differentiated.

Now we consider the propagation of the solution. For $k \in \{1, \cdots, K-1\}$ we say that the solution $(\tilde{\phi}(t, \mu), \tilde{P}(t, \mu))$ has the property $\mathcal{P}_k$ if:

(1) $\tilde{P}(t_k, \mu)$ and $\tilde{\phi}(t_k, \mu)$ are in $C^\infty(\bar{D}_r)$;

(2) Across the point p_k^2 for $k \geqq 2$, $\tilde{P}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k-1$, and $\tilde{\phi}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k-2$;

(3) Across the line segments L_k^2, $\tilde{P}(t, \mu)^{[j_1, j_2]}$ and $\tilde{\phi}(t, \mu)^{[j_1, j_2]}$ are continuous for $j_1 + j_2 \leqq k$;

(4) Across the point p_k^1, $\tilde{P}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k$ and $\tilde{\phi}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k-1$.

First we note that from our previous discussion, the solution has the property $\hat{\mathscr{P}}_1$. We now argue the theorem by induction. We suppose that for some positive integer k the solution has the property $\hat{\mathscr{P}}_j$ for $j \in \{1, \cdots, k\}$. Then it will follow that:

(I1) $\tilde{\phi}(t, \mu)$ and $\tilde{P}(t, \mu)$ are in $C^\infty(\bar{T}_k^1)$ and $C^\infty(\bar{T}_k^2)$;

(I2) Across L_k^1, $\tilde{P}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k$ and $\tilde{\phi}(t, \mu)^{[j_1, j_2]}$ is continuous for $j_1 + j_2 \leqq k-1$;

(I3) The solution $(\tilde{\phi}(t, \mu), \tilde{P}(t, \mu))$ has the property $\hat{\mathscr{P}}_{k+1}$ (for $k < K-1$).

Thus we now assume that the system has the property $\hat{\mathscr{P}}_j$ for $j \in \{1, \cdots, k\}$ where $k < K$, and we take steps to verify results (I1), (I2), and (I3). First we consider the system in T_k^1 and T_k^2. For $(t, \mu) \in T_k^1$ we have

$$\tilde{P}(t, \mu) = \tilde{P}(t_k, \mu) - \int_{t_k}^{\hat{t}_k(\mu)} B^T \tilde{\phi}_0(\tau) \tilde{\phi}^T(\tau, \mu)\, d\tau - \int_{\hat{t}_k(\mu)}^{t} B^T \tilde{\phi}_0(\tau) \tilde{\phi}^T(\tau, \mu)\, d\tau, \tag{6.15a}$$

$$\tilde{\phi}(t, \mu) = \tilde{\phi}_0(t+\mu) + \int_{t+\mu}^{t} \tilde{P}^T(\sigma, t+\mu-\sigma) B^T \tilde{\phi}_0(\sigma)\, d\sigma. \tag{6.15b}$$

For $(t, \mu) \in T_k^2$ we have

$$\tilde{P}(t, \mu) = \tilde{P}(t_k, \mu) - \int_{t_k}^{t} B^T \tilde{\phi}_0(\tau) \tilde{\phi}^T(\tau, \mu)\, d\tau, \tag{6.16a}$$

$$\tilde{\phi}(t, \mu) = \tilde{\phi}(t_k, \mu + t - t_k) + \int_{t_k}^{t} \tilde{P}^T(\sigma, t+\mu-\sigma) B^T \tilde{\phi}_0(\sigma)\, d\sigma. \tag{6.16b}$$

And for $t \in L_k^0$ we have

$$\begin{aligned} \tilde{\phi}_0(t) = \tilde{\phi}_0(t_k) - \int_{t_k}^{t} \Big\{ & [A_0^T - \tilde{P}_0^T(\tau) B^T] \tilde{\phi}_0(\tau) + A_1^T \tilde{\phi}_1(\tau) \\ & + \int_0^{\hat{\mu}_k(\tau)} A^T(\Theta) \tilde{\phi}(\tau, \Theta)\, d\Theta + \int_{\hat{\mu}_k(\tau)}^{r} A^T(\Theta) \tilde{\phi}(\tau, \Theta)\, d\Theta \Big\}\, d\tau, \end{aligned} \tag{6.17a}$$

$$\tilde{\phi}_1(t) = \tilde{\phi}_0(t_k, t - t_{k+1}) + \int_{t_k}^{t} \tilde{P}^T(\sigma, t+\mu-\sigma) B^T \tilde{\phi}_0(\sigma)\, d\sigma, \tag{6.17b}$$

$$\tilde{P}_0(t) = \tilde{P}_0(t_k) - \int_{t_k}^{t} B^T \tilde{\phi}_0(\tau) \tilde{\phi}_0^T(\tau)\, d\tau. \tag{6.17c}$$

By the inductive assumption we have that $\tilde{P}(t_k, \mu)$ and $\tilde{\phi}(t_k, \mu)$ are in $C^\infty(\bar{D}_r)$. Then by inspection we note that the system (6.15)-(6.17) is closed for the unknowns $\tilde{P}(t, \mu)$, $\tilde{\phi}(t, \mu)$, $\tilde{\phi}_0(t)$, $\tilde{P}_0(t)$, and $\tilde{\phi}_1(t)$. Repeated differentiation is possible, and the result (I1) is verified.

Next we consider smoothness across the line segment L_k^1. For $(t, \mu) \in T_k^2$ (6.16b) can be replaced by

$$\tilde{\phi}(t, \mu) = \tilde{\phi}_0(t+\mu) + \int_{t+\mu}^{t} \tilde{P}^T(\sigma, t+\mu-\sigma) B^T \tilde{\phi}_0(\sigma)\, d\sigma, \tag{6.18}$$

which is identical to (6.15b) except that here the integral crosses the line segment L_k^2. By the inductive assumption we are guaranteed that the solution has the property $\hat{\mathscr{P}}_k$; then $\tilde{\phi}_0(t)$ has derivatives of order $(k-1)$ continuous across t_k and $\tilde{P}(t, \mu)$ has derivatives of order k continuous across L_k^2. Thus the integral in (6.18) can be differentiated k times continuously for $(t, \mu) \in T_k^2$. Now as in (6.9) we develop recursive

formulas for derivatives of $\tilde{P}(t,\mu)$ and $\tilde{\phi}(t,\mu)$ in T_k^1 and T_k^2 and for $\tilde{\phi}_0(t)$, $\tilde{P}_0(t)$, $\tilde{\phi}_1(t)$, and $\tilde{P}_1(t)$ in L_k^0:

$$\tilde{\phi}_0(t)^{[1]} = (-A_0^T + \tilde{P}_0^T(t)B^T)\tilde{\phi}_0(t) - A_1\tilde{\phi}_1(t) + \int_0^{\hat{\mu}_k(t)} A^T(\Theta)\tilde{\phi}(t,\Theta)\,d\Theta$$

(6.19a)

$$+ \int_{\hat{\mu}_k(t)}^{r} A^T(\Theta)\tilde{\phi}(t,\Theta)\,d\Theta,$$

$$\tilde{P}_0(t)^{[1]} = -B^T\tilde{\phi}_0(t)\tilde{\phi}_0^T(t), \tag{6.19b}$$

$$\tilde{P}(t,\mu)^{[0,1]} = -B^T\tilde{\phi}_0(t)\tilde{\phi}^T(t,\mu), \tag{6.19c}$$

$$\tilde{P}(t,\mu)^{[1,0]} = \tilde{P}(t_k,\mu)^{[0,1]} - B^T\tilde{\phi}_0(t)\tilde{\phi}^T(t,\mu) + B^T\tilde{\phi}_0(t_k)\tilde{\phi}^T(t_k,\mu)$$

(6.19d)

$$+ \int_{t_k}^{t} B^T\tilde{\phi}_0(\tau)\tilde{\phi}^T(\tau,\mu)\,d\tau + \int_{t_k}^{t} B^T\tilde{\phi}_0(\tau)\tilde{\phi}^T(\tau)B\tilde{P}(\tau,\mu)\,d\tau,$$

$$\tilde{\phi}(t,\mu)^{[0,1]} = \tilde{\phi}_0(t+\mu^{[1]} - \tilde{P}_0^T(t+\mu)B^T\tilde{\phi}_0(t+\mu)$$

(6.19e)

$$+ \int_{t+\mu}^{t} \tilde{P}^T(t, t+\mu-\sigma)^{[0,1]}B^T\tilde{\phi}(\sigma)\,d\sigma,$$

$$\tilde{\phi}(t,\mu)^{[0,1]} = \tilde{\phi}(t,\mu)^{[0,1]} + \tilde{P}^T(t+\mu)B^T\tilde{\phi}_0(t), \tag{6.19f}$$

$$\tilde{\phi}_1(t)^{[1]} = \tilde{\phi}_0(t+r)^{[1]} - \tilde{P}_0^T(t+r)B^T\tilde{\phi}_0(t+r) + \int_{t+\mu}^{t} \tilde{P}^T(\sigma, t+r-\sigma)^{[0,1]}B^T\tilde{\phi}_0(\sigma)\,d\sigma$$

(6.19g)

$$+ \tilde{P}_1^T(t)B^T\tilde{\phi}_0(t),$$

$$\tilde{P}_1(t)^{[1]} = -B^T\tilde{\phi}_0(t)\tilde{\phi}_1^T(t). \tag{6.19h}$$

These expressions can be applied repeatedly to formulate derivatives on either side of L_k^1. The smoothness of the solution across t_k is given by the inductive hypothesis. In particular, we have that $\tilde{\phi}_0(t)$ has $(k-1)$ continuous derivatives across t_k and that $\tilde{P}_0(t)$ has k continuous derivatives across t_k. It then follows that $\tilde{\phi}(t,\mu)$ has $(k-1)$ continuous derivatives across L_k^1 and also that $\tilde{P}(t,\mu)$ has k continuous derivatives across L_k^1. This proves result (I2).

Next we consider the evaluation of these formulas at $t = t_{k+1}$. The system is closed with respect to these values except for the influence of previously analyzed values, which can be interpreted as forcing terms. First we consider the system (6.19) with the assumption that values corresponding to $t > t_{k+1}$ are known. We conclude from (6.19e)-(6.19g) and previous analysis of the solution for $t > t_{k+1}$ that $\tilde{\phi}(t,\mu)$ has continuous derivatives of order $(k-1)$ across p_{k+1}^2 and that $\tilde{\phi}_1(t)$ has continuous derivatives of order $(k-1)$ across t_{k+1}.

Next we consider the system (6.19a)-(6.19h) and assume that $\tilde{\phi}_1(t)$ is known. We conclude from (6.19a), (6.19e), (6.19f) and our previous analysis that $\tilde{\phi}_0(t)$ has continuous derivatives of order k across t_{k+1} and that $\tilde{\phi}(t,\mu)$ has continuous derivatives of order k across p_{k+1}^1.

Next we consider the system (6.19c)-(6.19f) and assume that $\tilde{\phi}_0(t)$ is known. By our previous analysis we conclude that $\tilde{\phi}(t,\mu)$ and $\tilde{P}(t,\mu)$ have continuous derivatives of order $(k+1)$ across L_{k+1}^2.

We complete our analysis at p_{k+1}^1 and p_{k+1}^2 by an additional examination of (6.19c), (6.19d). By our previous analysis we conclude that $\tilde{P}(t,\mu)$ has continuous derivatives of order k across p_{k+1}^2 and continuous derivatives of order $k+1$ across p_{k+1}^1. And finally we note that result (I1) guarantees that $\tilde{\phi}(t_{k+1},\mu)$ and $\tilde{P}(t_{k+1},\mu)$ are in $C^\infty(\bar{D}_r)$. This verifies result (I3) and completes the inductive argument. □

We conclude this section with a local formulation of the system (6.4)-(6.7) defined on a closed domain. Let α and β be positive real numbers such that

$$0 \leqq \beta < \alpha \leqq T.$$

As illustrated in Fig. 3, we define the sets

$$\begin{aligned} W_\alpha^\beta &= \{(t,\mu) \mid \beta \leqq t \leqq \alpha, 0 \leqq \mu \leqq r\}, \\ V_\alpha^\beta &= \{(t,\mu) \mid \beta \leqq t \leqq \alpha, 0 \leqq \mu \leqq r, t+\mu \leqq T\}, \\ Y_\alpha^\beta &= \{t \mid \beta \leqq t \leqq \alpha\}. \end{aligned} \tag{6.20}$$

Now let S_α^β denote the space of functions $\tilde{z}$ such that

$$\begin{aligned} &\tilde{z}: W_\alpha^\beta \to R^{(n+m)\times n}, \\ &\tilde{z}(t,\mu) \in C(V_\alpha^\beta), \\ &(t,\mu) \in W_\alpha^\beta \setminus V_\alpha^\beta \Rightarrow \tilde{z}(t,\mu) = 0. \end{aligned} \tag{6.21}$$

Then S_α^β has the structure of a Banach space with the norm

$$|\tilde{z}| = \sup |\tilde{z}(t,\mu)|_\infty, (t,\mu) \in V_\alpha^\beta. \tag{6.22}$$

The aim now is to represent the system (6.4)-(6.7) in this framework. First we make the identification

$$\tilde{z}(t,\mu) = \begin{pmatrix} \tilde{\phi}(t,\mu) \\ \tilde{P}(t,\mu) \end{pmatrix}, \tag{6.23}$$

whereby $\tilde{\phi}(t,\mu)$ and $\tilde{P}(t,\mu)$ are likewise extended into $W_\alpha^\beta \setminus V_\alpha^\beta$.

For $\tilde{z} \in S_\alpha^\beta$ we now define the mappings

$$\tilde{F}[\tilde{z}(\alpha,\mu)]: W_\alpha^\beta \to R^{(n+m)\times n}, \tag{6.24a}$$

$$\tilde{F}[\tilde{z}(\alpha,\mu)](t,\mu) = \begin{pmatrix} \tilde{F}_1[\tilde{z}(\alpha,\mu)](t,\mu) \\ \tilde{F}_2[\tilde{z}(\alpha,\mu)](t,\mu) \end{pmatrix}, \tag{6.24b}$$

$$\tilde{F}_1[\tilde{z}(\alpha,\mu)](t,\mu) = \begin{cases} \tilde{\phi}(\alpha, \mu-(\alpha-t)), & r \geqq \mu \geqq (\alpha-t), \\ \tilde{\phi}_0(\alpha), & (\alpha-t) > \mu \geqq 0, \end{cases} \tag{6.24c}$$

$$\tilde{F}_2[\tilde{z}(\alpha,\mu)](t,\mu) = \tilde{P}(\alpha,\mu), \tag{6.24d}$$

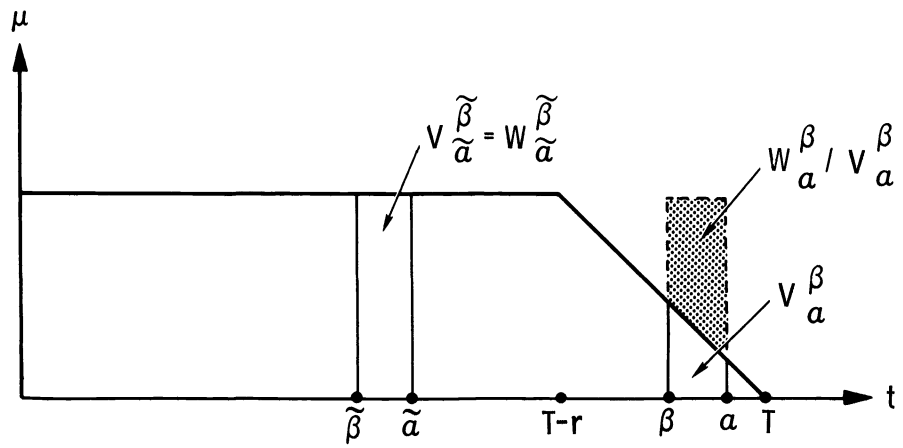


FIG. 3. *Local integral formulation.*

and

$$\tilde{G}[\tilde{z}(t,\mu)]: Y_\alpha^\beta \times W_\alpha^\beta \to R^{(n+m)\times n}, \tag{6.25a}$$

$$\tilde{G}[\tilde{z}(t,\mu)](\sigma,t,\mu) = \begin{pmatrix} \tilde{G}_1[\tilde{z}(t,\mu)](\sigma,t,\mu) \\ \tilde{G}_2[\tilde{z}(t,\mu)](\sigma,t,\mu) \end{pmatrix}, \tag{6.25b}$$

$$\tilde{G}_1[\tilde{z}(t,\mu)](\sigma,t,\mu) = \begin{cases} \{-[A_0^T - \tilde{P}_0^T(\sigma)B^T]\tilde{\phi}_0(\sigma) - \sum_{i=1}^{\nu} A_i^T \tilde{\phi}_i(\sigma) \\ -\int_0^r A^T(\Theta)\tilde{\phi}(\sigma,\Theta)\,d\Theta\}, & (t+\mu) \leqq \sigma \leqq \alpha, \\ \tilde{P}^T(\sigma, t+\mu-\sigma)B^T\tilde{\phi}_0(\sigma), & \beta \leqq \sigma < (t+\mu), \end{cases} \tag{6.25c}$$

$$\tilde{G}_2[\tilde{z}(t,\mu)](\sigma,t,\mu) = -B^T\tilde{\phi}_0(\sigma)\tilde{\phi}^T(\sigma,\mu). \tag{6.25d}$$

Using these definitions, we now give a local integral formulation that is similar to the system (6.15)-(6.17), which was derived for the case $\nu = 1$. Thus for $(t,\mu) \in W_\alpha^\beta$ we have

$$\tilde{z}(t,\mu) = \tilde{F}[\tilde{z}(\alpha,\mu)](t,\mu) + \int_\alpha^t \tilde{G}[\tilde{z}(t,\mu)](\sigma,t,\mu)\,d\sigma. \tag{6.26}$$

In particular, for $\alpha = T$ we have

$$\begin{aligned} &\tilde{z}(T,\mu) = \begin{pmatrix} \tilde{\phi}(T,\mu) \\ \tilde{P}(T,\mu) \end{pmatrix}, \\ &\tilde{\phi}(T,\mu) = \begin{cases} Q^{1/2}, & \mu = 0, \\ 0, & 0 < \mu \leqq r, \end{cases} \\ &\tilde{P}(T,\mu) = 0. \end{aligned} \tag{6.27}$$

This formulation of the equations is useful for both theoretical and practical purposes. In particular, we note that the structure is very similar to that of a parameterized system of ordinary differential equations (see, for example, [3]).

We can study the well posedness of the system as a functional equation in S_α^β. Thus with $\tilde{z}(\alpha,\mu)$ given we define the mapping $\tilde{T}$:

$$\begin{aligned} &\tilde{T}: S_\alpha^\beta \to S_\alpha^\beta, \\ &\tilde{T}\tilde{z}_1(t,\mu) = \tilde{F}[\tilde{z}(\alpha,\mu)](t,\mu) + \int_\alpha^t \tilde{G}[\tilde{z}_1(t,\mu)](\sigma,t,\mu)\,d\sigma. \end{aligned} \tag{6.28}$$

This definition leads to the estimate

$$|\tilde{T}\tilde{z}_1(t,\mu) - \tilde{T}\tilde{z}_2(t,\mu)| \leqq |\beta-\alpha|\tilde{K}|\tilde{z}_1(t,\mu) - z_2(t,\mu)|, \tag{6.29}$$

where $\tilde{K}$ is bounded in terms of γ and $\hat{\gamma}$ where we define

$$\gamma = \max\{n, |B|, |Q|, \max_i |A_i|, |A(\cdot)|_2\} \tag{6.30}$$

and we require

$$|z_1(t,\mu)| < \hat{\gamma}, \qquad |z_2(t,\mu)| < \hat{\gamma}. \tag{6.31}$$

Here $|\cdot|_2$ denotes the L_2 norm. Then for sufficiently small $|\beta-\alpha|$, $\tilde{T}$ defines a contraction mapping within some neighborhood of $z(t,\alpha)$ in S_α^β (see, for example, [22, p. 128]). Thus by the same arguments that are typically applied to systems of ordinary differential equations [3], [22], we have justification for the local existence and uniqueness of the solutions to the integral system (6.26). This in effect completes the uniqueness argument of Theorem 5.3.

As we shall demonstrate in the next section, this formulation is also convenient as a point of departure for the development of finite-difference approximations since the integrals in (6.26) can be replaced by quadratures very easily. Then estimates for local accuracy follow from Theorems 6.1 and 6.2, and estimates for stability likewise follow since the discretized system also yields a contraction mapping.

7. Finite-difference approximations. In this section we consider the development of finite-difference approximations for the optimal gain. As noted in § 6, the equations have the structure of a parameterized system of ordinary differential equations, and the same approximation techniques can be used (see, for example, [20]). Our intent here is not to provide an exhaustive analysis but rather to illustrate what is possible. We consider the general system in its integral formulation (6.4)-(6.7).

Without loss of generality we can assume as in Theorem 6.2 that $T = Kr$ for some positive integer K since the original problem can always be embedded in such a formulation. We focus on the local integral formulation (6.26) and consider discretizations of the set W_T^0 (cf. (6.20)). Thus for any positive integer N we define the grid parameter h and the grid G_h as follows:

$$\begin{aligned} &G_h = \{(t_i, \mu_j) \mid 0 \leqq i \leqq KN,\ 0 \leqq j \leqq N\}, \\ &h = r/N, \quad t_i = ih, \quad \mu_j = jh. \end{aligned} \tag{7.1}$$

Such a discretization is illustrated in Fig. 4. Conceptually, the limit $h \to 0$ corresponds to a sequence of finer and finer resolutions of W_T^0.

We now consider a set of vectors $\{\hat{z}_i\}_{i=0}^{NK}$ that have the form

$$\hat{z}_i = \begin{pmatrix} \hat{z}_i^{(0)} \\ \hat{z}_i^{(1)} \\ \hat{z}_i^{(2)} \\ \vdots \\ \hat{z}_i^{(N)} \end{pmatrix}, \qquad 0 \leqq i \leqq NK, \tag{7.2}$$

$$\hat{z}_i^{(j)} \in R^{(n+m)\times n}, \qquad 0 \leqq j \leqq N,$$

where $\hat{z}_i^{(j)}$ will be determined as an approximation to $\tilde{z}(ih, jh)$ on the grid G_h. Since we are interested in pointwise error estimates, the appropriate vector norm is defined

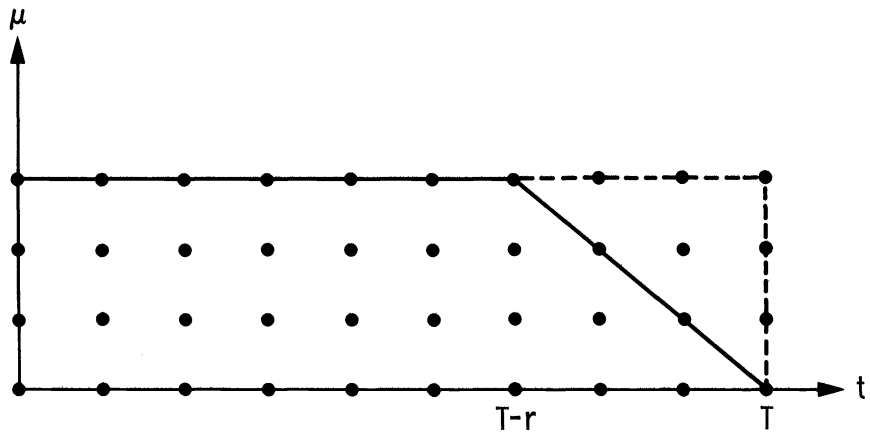


FIG. 4. *Discretization by point values.*

as

$$|\hat{z}_i| = \max_{(j)} |\hat{z}_i^{(j)}|_\infty. \tag{7.3}$$

To determine the $\hat{z}_i$'s, we introduce a one-step recursive definition that corresponds to a quadrature approximation of the integral formulation (6.26)-(6.27)

$$\begin{aligned} &\hat{z}_i = \hat{F}_h[\hat{z}_{i+1}] + h\hat{G}_h[\hat{z}_{i+1}], \qquad 0 \leqq i \leqq KN-1, \\ &\hat{z}_{KN} = ((Q^{1/2}, 0)^T, (0,0)^T, \cdots, (0,0)^T)^T. \end{aligned} \tag{7.4}$$

Now it is straightforward to develop notions of accuracy and stability that are analogous to what has been accomplished for systems of ordinary differential equations with prescribed initial conditions (cf. [20, p. 115]). Let γ be defined as in (6.30) and let $\tilde{\gamma}$ be some positive constant that bounds the solution:

$$\sup |\tilde{z}(t,\mu)|_\infty < \tilde{\gamma}, \qquad (t,\mu) \in W_T^0. \tag{7.5}$$

To study the accuracy of the approximation scheme we introduce the projection Λ_i defined for $i \in \{0, \cdots, NK\}$ as follows:

$$\begin{aligned} &\Lambda_i : S_T^0 \to R^{(n+m)\times n\times(N+1)}, \\ &\Lambda_i \tilde{z}(t,\mu) = \begin{pmatrix} \tilde{z}(t_i, 0) \\ \tilde{z}(t_i, h) \\ \tilde{z}(t_i, 2h) \\ \vdots \\ \tilde{z}(t_i, r) \end{pmatrix}. \end{aligned} \tag{7.6}$$

Then the approximation scheme (7.4) is said to have accuracy of order p, where p is some positive integer, if for some positive K_1, which depends only on γ, $\tilde{\gamma}$, and p, we have

$$\Lambda_i \tilde{F}[\tilde{z}(t,\mu)] - \hat{F}_h[\Lambda_i \tilde{z}(t,\mu)] = 0, \tag{7.7a}$$

$$\left| \Lambda_i \left(\int_{(i+1)h}^{ih} \tilde{G}[\tilde{z}(t,\mu)](\sigma, t, \mu)\, d\sigma \right) - h\hat{G}_h[\Lambda_{i+1}\tilde{z}(t,\mu)] \right| < K_1 h^{p+1}. \tag{7.7b}$$

As we noted at the end of the previous section, the stability of an approximation scheme of the form (7.4) is, in general, easy to verify since a discretization of (6.26) can also be analyzed as a contraction mapping (cf. (6.28)-(6.31)).

Now let the approximation scheme (7.4) be stable and have accuracy of order p. It is straightforward to show (cf. [20, p. 116]) that the discretization error can be bounded globally. More precisely, we can prove that there exists some positive K_2 that depends only on γ, $\tilde{\gamma}$, and p such that for sufficiently small h we have the estimate

$$\max_{0 \leqq i \leqq KN} |\Lambda_i \tilde{z}(t,\mu) - \hat{z}_i| < K_2 h^P. \tag{7.8}$$

This result completes our basic justification of the difference methods we employ since the parameter K_2 provides a modulus of convergence for the corresponding sequence of approximations.

To implement this theory requires approximate definitions for $\hat{F}_h[\cdot]$ and $\hat{G}_h[\cdot]$. Thus the integral equation (6.26) must be replaced by a quadrature scheme on the grid G_h, where the value of p in (7.7b) is determined by the order of the scheme and the smoothness of the solution $\tilde{z}(t,\mu)$. This latter condition is settled by an appeal to Theorems 6.1 and 6.2.

For example, we consider the general problem (4.1)-(4.2) with multiple delays. First we assume that the grid is chosen so that

$$r_i/h \in N, \qquad i \in \{1, 2, 3, \cdots, \nu\}. \tag{7.9}$$

This ensures that, for $i \in \{0, \cdots, NK-1\}$, the integral paths that determine

$$\int_{t_i}^{t_{i+1}} \tilde{G}[\tilde{z}(t, \mu)](\sigma, t, \mu)\, d\sigma$$

do not cross the line segments that define the piecewise-C^2 structure of the solution (see Theorem 6.1 and Fig. 1). This justifies a second-order method ($p=2$ in (7.7b)). We also can relax the restriction (7.9) to permit more general grids. Then locally the accuracy of the quadrature approximation is no better than first order when an integral path crosses one of the line segments ($p=1$ in (7.7b)); however, we can show that only a finite number of such errors can occur along each integral path. Thus, we have justification for a second-order method under very general conditions.

To illustrate the approach, we consider an implementation of the trapezoidal rule as a second-order predictor-corrector method (one forward-Euler predictor followed by two trapezoidal-rule correctors; cf. [20, p. 85]). We consider the following scalar system (4.1)-(4.2) with two delays:

$$\begin{aligned} &\dot{x}(t) = x(t) + x(t-r_1) + x(t-r_2) + u(t), \\ &r_1 = 0.2, \qquad r_2 = r = 1.0, \\ &Q = 1, \qquad T = 2.0. \end{aligned} \tag{7.10}$$

For any integer N let the approximation using the grid (7.1) be given as in (7.2) by

$$\hat{z}_i^N = \begin{pmatrix} \hat{\phi}_i^N \\ \hat{P}_i^N \end{pmatrix}, \qquad 0 \leqq i \leqq KN. \tag{7.11}$$

We consider approximations for the grids corresponding to

$$N \in \{10, 20, 25, 50\},$$

and also we consider as a reference discretization

$$N_* = 100.$$

According to (4.8)-(4.9) the gain at time t is determined by $P(t, s) = \tilde{P}(t, \mu)$ ($\mu = s - t$; cf. (6.1)). To study the convergence of approximations to $\tilde{P}(0, \mu)$, which determines the gain at $t=0$, we define the reference value

$$\hat{P}_* = \max_{0 \leqq j \leqq N_*} |\hat{P}_0^{N(j)}|.$$

Then for each N, a measure of the relative error is given by

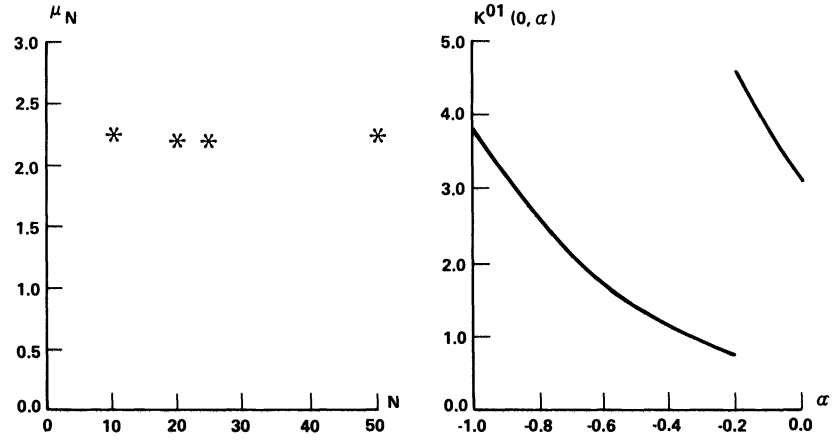
$$\varepsilon_N = \max_{0 \leqq j \leqq N} \frac{|\hat{P}_0^{N_*(jN_*/N)} - \hat{P}_0^{N(j)}|}{\hat{P}_*} \tag{7.12}$$

since (N^*/N) is always integral. In Fig. 5(a) we plot the associated quantity

$$\mu_N = \log \varepsilon_N / \log h \tag{7.13}$$

to verify the quadratic convergence of the approximations (cf. (7.8)). And also in Fig. 5(b) we plot the N_*-approximation to the integral component of the optimal gain at $t=0$ (cf. (4.7)-(4.9)):

$$K^{01}(0, \alpha) = \chi(-r_1, 0)(\alpha)\tilde{P}(0, \alpha + r_1) + \chi(-r_2, 0)(\alpha)\tilde{P}(0, \alpha + r_2). \tag{7.14}$$



(a) (b)

FIG. 5. *Example with two delays* (7.10): (a) *Quadratic convergence of approximations.* (b) *Integral component of optimal gain.*

A second example, taken from [17] and [2, p. 44], is based on a model for the fine tuning of the mach number in a cryogenic wind tunnel. The system, having a single delay ($\nu = 1$), has the form given by (4.1)-(4.2) with

$$A_0 = \begin{bmatrix} -a & 0 & 0 \\ 0 & -2b\omega & -\omega^2 \\ 0 & 1 & 0 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} 0 & 0 & ka \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{7.15}$$

$$B^T = (0, \omega^2, 0), \quad r = .33, \quad Q = \operatorname{diag}(10^4, 0, 0),$$

where the parameters $1/a$, ω^2, k, and b have the values 1.964, 36, $-.0117$, and .8, respectively. We choose for the final time

$$T = 6.6, \tag{7.16}$$

which comprises 20 delay intervals and, according to the results of [2], is sufficiently large to ensure that the values at $t = 0$ are nearly equal to the steady-state values.

As in the first example, we consider the convergence of the gain at $t = 0$. Since by (4.9) the integral component of the optimal gain has the form

$$K^{01}(0, \alpha) = \chi(-r, 0)(\alpha)\tilde{P}(0, \alpha + r)A_1, \tag{7.17}$$

only the first component of $\tilde{P}(0, \alpha + r)$ is relevant. Then for the same values of N and N_* as in the first example, we define

$$\hat{P}_* = \max_{0 \leqq j \leqq N_*} |e_1^T \hat{P}_0^{N_*(j)}|,$$

$$\varepsilon_N = \max_{0 \leqq j \leqq N} \frac{|e_1^T(\hat{P}_0^{N_*(jN_*/N)} - \hat{P}_0^{N(j)})|}{\hat{P}_*},$$

$$\mu_N = \log \varepsilon_N / \log h, \qquad e_1^T = (1, 0, 0).$$

And thus in Fig. 6(a) we plot the logarithmic error measure μ_N. In Fig. 6(b) we plot the N_*-approximation of the one nonzero entry of (7.17), which is given by $e_3^T K^{01}(0, \alpha)$ where $e_3^T = (0, 0, 1)$ (cf. Fig. 5.14 in [2]).

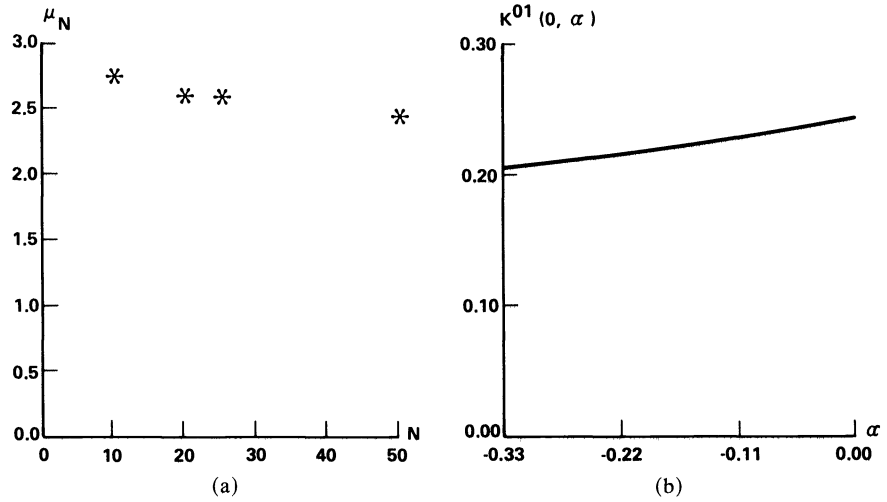


FIG. 6. *Example with one delay* (7.15): (a) *Quadratic convergence of approximations.* (b) *Integral component of optimal gain.*

In a similar fashion the results of Theorem 6.2 can be used to justify the existence of methods of arbitrarily high order for problems with a single delay and a smooth integral term. That is, the grid defined by (7.2) ensures that the relevant integral paths do not cross the lines that define the piecewise-C^∞ structure of the solution (cf. Fig. 4).

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.

[2] J. A. BURNS AND R. K. POWERS, *Factorization and reduction methods for optimal control of distributed parameter systems*, ICASE Report No. 85-88, NASA Langley Research Center, Hampton, VA, November 1985.

[3] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[4] R. CURTAIN AND A. J. PRITCHARD, *The infinite-dimensional Riccati equation for systems defined by evolution operators*, SIAM J. Control Optim., 14 (1976), pp. 951-983.

[5] M. C. DELFOUR, *The linear quadratic optimal control problem for hereditary differential systems: theory and numerical solution*, Appl. Math. Optim., 3 (1977), pp. 101-162.

[6] M. C. DELFOUR AND A. MANITIUS, *The structural operator F and its role in the theory of retarded system* I, J. Math. Anal. Appl., 73 (1980), pp. 466-490.

[7] D. H. ELLER, J. K. AGGARWAL, AND H. T. BANKS, *Optimal control of linear time-delay systems*, IEEE Trans. Automat. Control, 14 (1969), pp. 678-687.

[8] J. S. GIBSON, *Linear-quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95-139.

[9] ———, *The Riccati integral equations for optimal control problems on Hilbert space*, SIAM J. Control Optim., 17 (1979), pp. 537-565.

[10] I. C. GOHBERG AND I. KOLTRACHT, *Numerical solution of integral equations, fast algorithms and Krein-Sobolev equations*, Numer. Math., 47 (1985), pp. 237-288.

[11] I. C. GOHBERG AND M. G. KREIN, *Theory and Applications of Volterra Operators in Hilbert Space*, American Mathematical Society, Providence, RI, 1970.

[12] J. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[13] K. ITO, *Strong convergence and convergence rates of approximating solutions for algebraic Riccati equations in Hilbert space*, in Proc. 3rd Internat. Conference on Distributed Parameter Systems, Vorau, Styria, July 1986, Springer-Verlag, New York, 1987.

[14] K. ITO AND R. POWERS, *Chandrasekhar equations for infinite dimensional systems*, SIAM J. Control Optim., 25 (1987), pp. 596-611.

[15] K. ITO AND R. TEGLAS, *Legendre-Tau approximations for functional differential equations*, SIAM J. Control Optim., 24 (1986), pp. 737-760.

[16] T. KAILATH, B. LEVY, L. LJUNG, AND M. MORF, *The factorization and representation of operators in the algebra generated by Toeplitz operators*, SIAM J. Appl. Math., 37 (1979), pp. 467-484.

[17] F. KAPPEL AND D. SALAMON, *Spline approximation for retarded systems and the Riccati equation*, SIAM J. Control Optim., 25 (1987), pp. 1082-1117.

[18] M. KROLLER, *Numerische Approximation des Linear Quadratischen Optimierungsproblems Bei Paragolischen Differentialgleichugen*, Ph.D. dissertation, University of Graz, Graz, Austria, September 1986.

[19] K. KUNISCH, *Approximation schemes for the linear-quadratic optimal control problem associated with delay equations*, SIAM J. Control Optim., 20 (1982), pp. 506-540.

[20] J. D. LAMBERT, *Computational Methods in Ordinary Differential Equations*, John Wiley, New York, 1973.

[21] P. LINZ, *Analytical and Numerical Methods for Volterra Equations*, Siam Studies in Applied Mathematics 7, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1985.

[22] ———, *Theoretical Numerical Analysis*, John Wiley, New York, 1979.

[23] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Dunod, 1969.

[24] A. MANITIUS, *Optimal control of linear time-lag processes with quadratic performance indexes*, in Proc. 4th IFAC Congress, Warsaw, Poland, 1969, pp. 16-28.

[25] A. MCNABB AND A. SCHUMITZKY, *Factorization of operators* II: *a nonlinear Volterra method for numerical solution of linear Fredholm equations*, J. Comput. System Sci., 4 (1970), pp. 103-128.

[26] M. MILMAN, *Approximating the linear quadratic optimal control law for hereditary systems with delays in the control*, SIAM J. Control Optim., 26 (1988), pp. 291-320.

[27] ———, *An extension of the special factorization with applications to Wiener-Hopf equations*, J. Math. Anal. Appl., 110 (1985), pp. 303-322.

[28] ———, *Special factorization and Riccati integral equations*, J. Math. Anal. Appl., 100 (1984), pp. 155-187.

[29] M. MILMAN, J. FOSTER, AND A. SCHUMITZKY, *Optimal feedback control of infinite dimensional linear systems with applications to hereditary problems*, J. Math. Anal. Appl., 119 (1986), pp. 259-281.

[30] M. MILMAN AND R. E. SCHEID, *Factorization and the synthesis of optimal feedback kernels for differential-delay systems*, in Proc. 26th Conference on Decision and Control, Los Angeles, CA, December 1987.

[31] ———, *Numerical methods for finite-time quadratic control of infinite-dimensional systems*, in Proc. 4th IFAC Symposium on Control of Distributed Parameter Systems, Los Angeles, CA, July 1986.

[32] M. MILMAN AND A. SCHUMITZKY, *On a class of operators on Hilbert space with applications to factorization and systems theory*, J. Math. Anal. Appl., 99 (1984), pp. 494-512.

[33] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[34] A. C. ZAANEN, *Linear Analysis*, Interscience Publishers Inc., North-Holland, New York, 1953.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 1103-1112, September 1990

004

A CHARACTERIZATION OF FEEDBACK EQUIVALENCE*

J. M. GRACIA†, I. DE HOYOS†, AND I. ZABALLA†

Abstract. This paper provides a new characterization of feedback equivalence that can be applied to controllable and noncontrollable matrix pairs (A, B). This result is based on a generalization of a theorem of Rosenbrock describing the closed-loop invariant polynomials that are attainable by applying state feedback to a given system $\dot{x} = Ax + Bu$.

Key words. feedback equivalence, state feedback, invariant factors assignment, controllability indices

AMS(MOS) subject classifications. 93B10, 93B25, 15A21

1. Introduction and notation. Let $(A, B) \in \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times m}$, where $\mathbb{C}$ is the complex field. Consider the ordinary state-space model

$$\dot{x}(t) = Ax(t) + Bu(t). \tag{1}$$

Let us review the definition of the state feedback group (see, e.g., [6], [3], [9, p. 118], [8, p. 122]). We consider three types of elementary transformation on the system (1): (i) change of basis in the state space $x = Pz$, with P a nonsingular $n \times n$ matrix; (ii) change of basis in the input space $u = Qv$, with Q a nonsingular $m \times m$ matrix; (iii) state feedback $u = Fx + v$. These operations transform the matrix pair (A, B) as follows:

$$(A, B) \to (P^{-1}AP, P^{-1}B), \tag{1.1}$$

$$(A, B) \to (A, BQ), \tag{1.2}$$

$$(A, B) \to (A + BF, B). \tag{1.3}$$

The transformation group generated by (1.1)-(1.3) can be conveniently represented in the following way. Let $H(n, m)$ denote the group of all nonsingular $(n + m) \times (n + m)$ matrices of the form

$$\begin{bmatrix} P & 0 \\ F & Q \end{bmatrix}$$

with P $n \times n$, F $m \times n$, Q $m \times m$. We refer to $H(n, m)$ as the *state feedback group.* Define a right group action of $H(n, m)$ on $\mathbb{C}^{n\times n} \times \mathbb{C}^{n\times m}$ by

$$\left((A, B), \begin{bmatrix} P & 0 \\ F & Q \end{bmatrix}\right) \to (P^{-1}AP + P^{-1}BF, P^{-1}BQ). \tag{2}$$

The transformations (1.1)-(1.3) correspond to the special cases of (2), where $F = 0$ and $Q = I$, $P = I$ and $F = 0$, $P = I$ and $Q = I$, respectively.

This action yields an equivalence relation on the set $\mathbb{C}^{n\times n} \times \mathbb{C}^{n\times m}$ called the *feedback equivalence.* It is of interest to know when two matrix pairs (A_1, B_1) and (A_2, B_2) are feedback equivalent, i.e., belong to the same orbit. A complete system of invariants for this equivalence relation is given by the controllability indices and the

* Received by the editors August 10, 1988; accepted for publication (in revised form) October 25, 1989.
† Departamento de Matemáticas, Universidad de País Vasco, Facultad de Farmacia, Apartado 450, E-01080 Vitoria-Gasteiz, Spain.

invariant polynomials of the pair (A, B) [10, Thm. 2.12] where the invariant polynomials of a pair (A, B) are defined as the invariant factors of the $n \times (n+m)$ polynomial matrix

$$[\lambda I - A, -B].$$

Our aim in this paper is to provide a new characterization of the feedback equivalence. We emphasize the role played by the set (*feedback set* of (A, B))

$$\{A+BF \mid F \in \mathbb{C}^{m\times n}\},$$

associated with (A, B). This set appears significantly in the study of some problems related to a pair (A, B). So, if $\mathrm{Inv}(A, B)$ denotes the set of all (A, B)-invariant subspaces, and $\mathrm{Inv}(M)$ denotes the lattice of M-invariant subspaces for a square matrix M, then

$$\mathrm{Inv}(A, B) = \bigcup_{F\in\mathbb{C}^{m\times n}} \mathrm{Inv}(A+BF)$$

[9, § 4.2]. A complex number λ_0 is said to be an *eigenvalue* (or *incontrollable pole*) of (A, B) if there exists a nonzero vector $x \in \mathrm{Ker}\, B^T$ such that

$$A^T x = \lambda_0 x.$$

Here superscript T stands for transpose. From the spectral assignment theorem [9, p. 50] we can prove that

$$\sigma(A, B) = \bigcap_{F\in\mathbb{C}^{m\times n}} \sigma(A+BF),$$

where $\sigma(A, B)$ denotes the set of eigenvalues of (A, B), and $\sigma(M)$ denotes the set of eigenvalues of a square matrix M.

Let us now consider another set of matrices (*extension set* of (A, B))

$$\left\{\begin{bmatrix} A & B \\ C & D \end{bmatrix} \mid C \in \mathbb{C}^{m\times n},\ D \in \mathbb{C}^{m\times m}\right\}$$

for a given pair $(A, B) \in \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times m}$. It turns out that for each of the above properties involving (A, B) and its feedback set there is one involving (A, B) and its extension set. Namely,

$$\sigma(A, B) = \bigcap_{\substack{C\in\mathbb{C}^{m\times n}\\ D\in\mathbb{C}^{m\times m}}} \sigma\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right), \tag{3}$$

$$\mathrm{Inv}(A, B) = P\left[\bigcup_{\substack{C\in\mathbb{C}^{m\times n}\\ D\in\mathbb{C}^{m\times m}}} \mathrm{Inv}\left(\begin{bmatrix} A & C \\ B & D \end{bmatrix}\right)\right], \tag{4}$$

where $P: \mathbb{C}^{n+m} \to \mathbb{C}^n$ is the projector defined by

$$P((x_1, \cdots, x_{n+m})^T) := (x_1, \cdots, x_n)^T,$$

and

$$P\{\mathcal{V}_1, \mathcal{V}_2, \cdots\} = \{P\mathcal{V}_1, P\mathcal{V}_2, \cdots\}$$

if $\mathcal{V}_1, \mathcal{V}_2, \cdots$ are subspaces of $\mathbb{C}^{n+m}$, [5, Thm. 6.1.1, p. 190].

This extraordinary and interesting parallelism between the feedback and the extension sets of a given pair does not end with these properties. Theorem 5.1 of [10] and Theorem 2.6 of [11] give characterizations of all possible invariant polynomials

of the matrices belonging to the feedback and extension sets of (A, B), respectively, and both of them are very similar. Actually, (3) is an immediate consequence of Corollary IV of Theorem 5.1 of [10].

On the other hand, in [5] and [10] it has been proved that two pairs of matrices (A_1, B_1), $(A_2, B_2) \in \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times m}$ are feedback equivalent if and only if there exists $T \in H(n, m)$ such that for any pair $(C_1, D_1) \in \mathbb{C}^{m\times n} \times \mathbb{C}^{m\times m}$ there is a pair $(C_2, D_2) \in \mathbb{C}^{m\times n} \times \mathbb{C}^{m\times m}$ satisfying

$$T\begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} T^{-1} = \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix}.$$

As far as we know there is no characterization of the feedback equivalence in terms of the feedback sets of (A_1, B_1) and (A_2, B_2), and this paper is devoted to closing this gap.

Although we have been considering matrices of complex numbers, the previous and the following results remain valid for any *arbitrary* field $\mathbb{F}$, and in the sequel we will assume this more general setting. Thus we will use the following notation. If $M \in \mathbb{F}^{n\times n}$, the greatest common divisor of all minors of order k of the polynomial matrix $\lambda I - M$ is called the *kth determinantal divisor* of M and is denoted by $D_k(M)$, $(k = 1, 2, \cdots, n)$. $\mathbb{F}[\lambda]$ will be the ring of polynomials in one variable λ with coefficients in $\mathbb{F}$; the degree of $\alpha \in \mathbb{F}[\lambda]$ is denoted by $d(\alpha)$, and $|$ means "divides."

If $x = (x_1, \cdots, x_m)$ and $y = (y_1, \cdots, y_m)$ are two *nonincreasing* m-tuples of integers, x is said to be *majorized* by y, and we write $x \prec y$, if

$$\sum_{i=1}^{k} x_i \leqq \sum_{i=1}^{k} y_i \quad \text{for } k = 1, 2, \cdots, m-1,$$

$$\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i.$$

If $(y_1, \cdots, y_n)$ and $(x_1, \cdots, x_m)$ are two tuples of integers such that $y_1 \leqq \cdots \leqq y_n$, $x_1 \leqq \cdots \leqq x_m$, then we call the *union* of these tuples the finite sequence $z_1 \leqq z_2 \leqq \cdots \leqq z_{n+m}$, formed by all the components $y_1, \cdots, y_n, x_1, \cdots, x_m$ rearranged in nondecreasing order. We denote it by

$$(z_1, \cdots, z_{n+m}) = (y_1, \cdots, y_n) \cup (x_1, \cdots, x_m).$$

For example, if $y = (1, 2, 3)$ and $x = (2, 2)$, then $y \cup x = (1, 2, 2, 2, 3)$.

The organization of this paper is as follows. In § 2 we present the main theorem, Theorem 1. Then in § 3 we give some results that are needed for its proof, namely, a generalization of Rosenbrock's theorem on assignment of invariant polynomials by state feedback to the general case of noncontrollable systems (Theorem 2), and Lemma 4, which provides a method of constructing some polynomials that interlace some given polynomials and of satisfying a prescribed degree condition. In this way we solve an inverse problem for polynomials. Finally in § 4 we give the proofs.

2. Main result.

THEOREM 1 (a criterion for feedback equivalence). *Let* (A_1, B_1), $(A_2, B_2) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$ *be two matrix pairs. The two following statements are equivalent*:

(1) (A_1, B_1) *is feedback equivalent to* (A_2, B_2).

(2) *For each matrix* $F_1 \in \mathbb{F}^{m\times n}$ *there exists a matrix* $F_2 \in \mathbb{F}^{m\times n}$ *such that* $A_1 + B_1F_1$ *and* $A_2 + B_2F_2$ *are similar, and conversely, for each* $F_2 \in \mathbb{F}^{m\times n}$ *there exists an* $F_1 \in \mathbb{F}^{m\times n}$ *such that* $A_2 + B_2F_2$ *and* $A_1 + B_1F_1$ *are similar.*

The proof of this theorem is in § 4.

Remark. Condition (2) in Theorem 1 would not be sufficient if only one of the conditions in it is true, i.e., there exist matrix pairs (A_1, B_1), $(A_2, B_2) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$ that are not feedback equivalent and such that the following conditions hold:

(a) For each $F_1 \in \mathbb{F}^{m\times n}$ there exists an $F_2 \in \mathbb{F}^{m\times n}$ such that $A_1 + B_1F_1$ and $A_2 + B_2F_2$ are similar; and

(b) There exists an $F_2 \in \mathbb{F}^{m\times n}$ such that for all $F_1 \in \mathbb{F}^{m\times n}$, $A_2 + B_2F_2$ is not similar to $A_1 + B_1F_1$.

Proof. Let us suppose $m \leqq n$. Let $A_1 = 0$, $A_2 = 0$, be matrices of $\mathbb{F}^{n\times n}$. Let $B_1 = 0$, $B_2 = \binom{I_m}{0} \in \mathbb{F}^{n\times m}$, where I_m denotes the $m \times m$ identity matrix. It is clear that (A_1, B_1) and (A_2, B_2) are not feedback equivalent. On the other hand, for each $F_1 \in \mathbb{F}^{m\times n}$, $A_1 + B_1F_1 = 0$, there exists $F_2 = 0 \in \mathbb{F}^{m\times n}$ such that $A_2 + B_2F_2 = 0$, and so $A_1 + B_1F_1$ and $A_2 + B_2F_2$ are similar. Finally, it suffices to take $F_2 = (I_m\ 0) \in \mathbb{F}^{m\times n}$ in order to see that

$$A_2 + B_2F_2 = \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix}$$

and for all $F_1 \in \mathbb{F}^{m\times n}$ we have that $A_2 + B_2F_2$ is not similar to $A_1 + B_1F_1$, because $A_1 + B_1F_1 = 0$.

Now we explore a possible system-theoretic implication of Theorem 1. If $(A, B) \in \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times m}$ is not controllable we can define that (A, B) is *stabilizable* if and only if there exists a *gain matrix feedback* $F \in \mathbb{C}^{m\times n}$ such that the system $\dot{x} = (A + BF)x$ is asymptotically stable, i.e., each one of its solutions $x(t) \to 0$ when $t \to \infty$; a weaker property is that $x(t)$ is bounded when $t \to \infty$. It is well known that this can be characterized in terms of the real parts of the eigenvalues and of the size of Jordan blocks in the Jordan canonical form of $A + BF$ [2, Thm. 5.2, p. 178], [4, p. 398]. So, to assess the stabilizability of a system $\dot{x} = Ax + Bu$ by state feedback it is sufficient to know the *Jordan part* of A in the Brunovsky canonical form of the pair (A, B) [5, Thm. 6.2.5, p. 196].

If system (1) $\dot{x} = A_1x + B_1u$ is feedback equivalent to system (2) $\dot{x} = A_2x + B_2u$, then for each state feedback $u = F_1x$ that we can perform on (1), there exists a state feedback $u = F_2x$ on (2) such that the solutions of the systems $\dot{x} = (A_1 + B_1F_1)x$ and $\dot{x} = (A_2 + B_2F_2)x$ have the same asymptotic behavior when $t \to \infty$.

3. State feedback. The theorem of Rosenbrock [7, Thm. 4.2, Cor. 1, pp. 190–192], [1, Thm. 4.4, p. 278] is an important result that describes precisely the invariant polynomials that can be assigned by performing a state feedback on a controllable system.

Rosenbrock's theorem can be seen as a consequence of the following theorem, which states a characterization of the possible invariant polynomials to be assigned by state feedback on *noncontrollable* systems. (We recall that a pair (A, B) is determined up to state feedback equivalence by its invariant polynomials and controllability indices). Although this theorem was proved in [11] we are giving it for the reader's convenience.

THEOREM 2 (invariant polynomials assignment by feedback) [11, Thm. 2.6]. *Let* $(A, B) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$. *Let* $\alpha_1 | \cdots | \alpha_n$ *be the invariant polynomials of* (A, B) *and* $k_1 \geqq \cdots \geqq k_m \geqq 0$ *its controllability indices. Let* $\gamma_1 | \cdots | \gamma_n$ *be* n *monic polynomials of* $\mathbb{F}[\lambda]$. *Then there exists a matrix* $F \in \mathbb{F}^{m\times n}$ *such that* $A + BF$ *has* $\gamma_1, \cdots, \gamma_n$ *as invariant polynomials if and only if*

$$\gamma_{i-m} | \alpha_i | \gamma_i, \qquad i = 1, \cdots, n, \tag{5}$$

$$(k_1, \cdots, k_m) \prec (d(\sigma_m), \cdots, d(\sigma_1)), \tag{6}$$

where

$$\sigma_j := \frac{\beta^j}{\beta^{j-1}}, \qquad \beta^j := \beta_1^j \cdots \beta_{n+j}^j,$$

$$\beta_i^j := \text{l.c.m.}\,(\alpha_{i-j}, \gamma_{i-m}), \qquad i = 1, \cdots, n+j, \quad j = 0, 1, \cdots, m,$$

and we agree that $\alpha_i = \gamma_i = 1$ *for* $i < 1$.

We can get Rosenbrock's theorem by setting $\alpha_i = 1$, $i = 1, \cdots, n$.

For the following theorem we need some additional concepts and notation. The controllability matrix $S(A, B)$ of $(A, B) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$ is the $n \times nm$ matrix $[B, AB, \cdots, A^{n-1}B]$. If $s = \text{rank}\, S(A, B)$, then $s = k_1 + \cdots + k_m$, where $k_1, \cdots, k_m$ are the controllability indices of (A, B). If $\alpha_1 | \cdots | \alpha_n$ are the invariant polynomials of (A, B), then $s = n - d(\prod_{i=1}^n \alpha_i)$ (see [10] or [5, Thm. 6.2.5, p. 196]).

THEOREM 3. *Let* $(A, B) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$. *Let* $\alpha_1 | \cdots | \alpha_n$ *be the invariant polynomials of* (A, B). *Then, we have for* $k = 1, \cdots, n$,

$$\prod_{j=1}^{k} \alpha_j = \text{g.c.d.}\,\{D_k(A + BF) \mid F \in \mathbb{F}^{m\times n}\},$$

where $D_k(M)$ *denotes the kth determinantal divisor of the matrix* M.

The proof of this theorem is in § 4.

As we said in the Introduction, in the following lemma we solve an inverse problem for polynomials. The relation between this lemma and Theorem 2 can be seen immediately.

LEMMA 4. *Let* m, n, s *be positive integers with* $n \geqq s$. *Let* $\alpha_1 | \cdots | \alpha_n$ *be monic polynomials of* $\mathbb{F}[\lambda]$ *such that* $\sum_{i=1}^n d(\alpha_i) = n - s$. *Let* $(x_1, \cdots, x_m)$ *be an* m*-tuple of integers such that* $0 \leqq x_1 \leqq \cdots \leqq x_m$ *and* $\sum_{j=1}^m x_j = s$. *Then there exist* n *monic polynomials* $\gamma_1 | \cdots | \gamma_n$ *of* $\mathbb{F}[\lambda]$ *such that*

$$\gamma_{i-m} | \alpha_i | \gamma_i \qquad (i = 1, \cdots, n), \tag{7}$$

$$d(\sigma_j) = x_j \qquad (j = 1, \cdots, m), \tag{8}$$

where

$$\sigma_j := \frac{\beta^j}{\beta^{j-1}}, \qquad \beta^j := \beta_1^j \cdots \beta_{n+j}^j,$$

$$\beta_i^j := \text{l.c.m.}\,(\alpha_{i-j}, \gamma_{i-m}) \quad \text{for } i = 1, \cdots, n+j, \quad j = 0, 1, \cdots, m;$$

and we agree that $\alpha_i = \gamma_i = 1$ *for* $i < 1$.

The proof of Lemma 4 will be given in § 4.

Finally, the following proposition shows the almost evident fact that two m-tuples of numbers are equal if they have the same set of upper bounds.

PROPOSITION 5. *Let* m, s *be positive integers. Let* $(k_1^1, \cdots, k_m^1)$ *and* $(k_1^2, \cdots, k_m^2)$ *be two* m*-tuples of integers with* $k_1^1 \geqq \cdots \geqq k_m^1 \geqq 0$, $k_1^2 \geqq \cdots \geqq k_m^2 \geqq 0$, $\sum_{i=1}^m k_i^1 = \sum_{i=1}^m k_i^2 = s$, *such that for all* m*-tuples of integers* $x_m \geqq \cdots \geqq x_1 \geqq 0$, *with* $\sum_{i=1}^m x_i = s$, *we have that* $(k_1^1, \cdots, k_m^1) < (x_m, \cdots, x_1)$ *if and only if* $(k_1^2, \cdots, k_m^2) < (x_m, \cdots, x_1)$. *Then,*

$$(k_1^1, \cdots, k_m^1) = (k_1^2, \cdots, k_m^2).$$

4. Proofs.

Proof of Theorem 3. Let $F \in \mathbb{F}^{m\times n}$ be any matrix. If $\gamma_1 | \cdots | \gamma_n$ are the invariant polynomials of $A + BF$, applying (5) we have

$$\prod_{j=1}^{k} \alpha_j \Big| \prod_{j=1}^{k} \gamma_j, \qquad k = 1, \cdots, n. \tag{9}$$

Let us define the following n polynomials:

$$\gamma_i' := \alpha_i, \quad 1 \leqq i \leqq n-1, \qquad \gamma_n' := \alpha_n \mu,$$

where $\mu \in \mathbb{F}[\lambda]$ is any polynomial of degree s, where $s = \operatorname{rank} S(A, B)$. These polynomials satisfy conditions (5) and (6). In fact, (5) holds trivially. Now, we will compute the polynomials β_i^j.

For $j < m$, as $i \leqq n+j$, we have $i < n+m$; thus $i-m < n$ and then $\gamma_{i-m} = \alpha_{i-m}$. Moreover, for $j < m$, $i-m < i-j$ and therefore $\alpha_{i-m} | \alpha_{i-j}$; that is to say $\gamma_{i-m} | \alpha_{i-j}$. Hence, for $j < m$ we have

$$\beta_i^j = \text{l.c.m.}\,(\alpha_{i-j}, \gamma_{i-m}) = \alpha_{i-j}, \qquad 1 \leqq i \leqq n+j.$$

From (5) we conclude that $\alpha_{i-m} | \gamma_{i-m}$, and so

$$\beta_i^m = \text{l.c.m.}\,(\alpha_{i-m}, \gamma_{i-m}) = \gamma_{i-m}, \qquad 1 \leqq i \leqq n+m.$$

Consequently, for $j = 1, \cdots, m-1$

$$\sigma_j = \frac{\beta_1^j \cdots \beta_{n+j}^j}{\beta_1^{j-1} \cdots \beta_{n+j-1}^{j-1}} = \frac{\alpha_{1-j} \cdots \alpha_n}{\alpha_{2-j} \cdots \alpha_n} = 1,$$

$$\sigma_m = \frac{\beta_1^m \cdots \beta_{n+m}^m}{\beta_1^{m-1} \cdots \beta_{n+m-1}^{m-1}} = \frac{\gamma_1 \cdots \gamma_n}{\alpha_1 \cdots \alpha_n} = \mu.$$

Since $d(\mu) = s = \sum_{j=1}^m k_j$, it is clear that (6) is satisfied. Therefore, by applying Theorem 2, there exists $F \in \mathbb{F}^{m \times n}$ such that $\alpha_1, \cdots, \alpha_n \mu$ are the invariant polynomials of $A + BF$. Thus, for this F, we have

$$\begin{aligned} D_k(A+BF) &= \prod_{j=1}^{k} \alpha_j, \qquad k = 1, \cdots, n-1, \\ D_n(A+BF) &= \left(\prod_{j=1}^{n} \alpha_j \right) \mu. \end{aligned} \tag{10}$$

Now, whichever the underlying field $\mathbb{F}$, we can always find two polynomials $\mu_1, \mu_2 \in \mathbb{F}[\lambda]$ such that the g.c.d. is $(\mu_1, \mu_2) = 1$. And for each one of these polynomials there exists a matrix $F_i \in \mathbb{F}^{m \times n}$ such that $A + BF_i$ has $\alpha_1, \cdots, \alpha_n \mu_i$ as its invariant polynomials, $(i = 1, 2)$. Consequently,

$$\text{g.c.d.}\,(D_n(A+BF_1), D_n(A+BF_2)) = \prod_{j=1}^{n} \alpha_j. \tag{11}$$

From (9)–(11), we conclude that

$$\prod_{j=1}^{k} \alpha_j = \text{g.c.d.}\,\{D_k(A+BF) \mid F \in \mathbb{F}^{m \times n}\}, \qquad k = 1, \cdots, n$$

and the theorem follows. □

Proof of Lemma 4. The lemma will be proved if we can find $n+m$ monic polynomials $\psi_1 | \cdots | \psi_{n+m}$ such that $\psi_i = 1$, $1 \leqq i \leqq m$, with the following properties:

$$\psi_i | \alpha_i | \psi_{i+m} \qquad (i = 1, \cdots, n), \tag{12}$$

$$d(\sigma_j) = x_j \qquad (j = 1, \cdots, m), \tag{13}$$

where

$$\sigma_j = \frac{\beta^j}{\beta^{j-1}}, \qquad \beta^j = \beta_1^j \cdots \beta_{n+j}^j,$$

$$\beta_i^j = \text{l.c.m.}\,(\alpha_{i-j}, \psi_i) \quad \text{for } i = 1, \cdots, n+j, \qquad j = 0, 1, \cdots, m.$$

In fact, if we find these polynomials and we set

$$\gamma_i := \psi_{m+i} \qquad (i = 1, \cdots, n),$$

then $\gamma_1, \cdots, \gamma_n$ will satisfy (7) and (8).

First suppose that there exists an irreducible polynomial $\phi_0 \in \mathbb{F}[\lambda]$ such that $\alpha_i = \phi_0^{y_i}$ for $i = 1, \cdots, n$. Let d_0 be the degree of ϕ_0. Thus, $(y_1, \cdots, y_n)$ is an n-tuple of integers such that $0 \leqq y_1 \leqq \cdots \leqq y_n$ and $d_0 \sum_{i=1}^{n} y_i = n - s$.

Let q_j and r_j be the quotient and the remainder obtained from the Euclidean division of x_j by d_0:

$$x_j = q_j d_0 + r_j, \quad 0 \leqq r_j < d_0 \quad (1 \leqq j \leqq m).$$

We define

$$\begin{aligned}(z_1, \cdots, z_{n+m}) &:= (y_1, \cdots, y_n) \cup (q_1, \cdots, q_m) \\ &= (y_1, \cdots, y_{g_1}, q_1, y_{g_1+1}, \cdots, y_{g_2}, q_2, y_{g_2+1}, \cdots, \\ &\quad y_{g_m}, q_m, y_{g_m+1}, \cdots, y_n).\end{aligned}$$

So, for $t = 1, \cdots, m$ and agreeing that $g_0 := 0$ and $g_{m+1} := m$, we have that

$$\begin{aligned}z_i &= y_{i-t+1} \quad \text{for } i = g_{t-1} + t, \cdots, g_t + t - 1, \\ z_i &= q_t \qquad \text{for } i = g_t + t,\end{aligned}$$

and $z_1 \leqq z_2 \leqq \cdots \leqq z_{n+m}$.

Since $\sum_{i=1}^{n} y_i \leqq n - s$ and $\sum_{i=1}^{m} q_i \leqq s$, we have that $y_1 = \cdots = y_s = 0$ and $q_1 = \cdots = q_{m-s} = 0$. Therefore, $z_1 = \cdots = z_m = 0$.

Now let $k \in \{1, \cdots, n\}$. Then there exists $t \in \{1, \cdots, m+1\}$ such that $k \in \{g_{t-1}+1, \cdots, g_t\}$ and by the definition of z_i it follows that

$$y_k = z_{k+t-1}. \tag{14}$$

From (14), we get

$$z_i \leqq y_i \leqq z_{i+m}, \qquad 1 \leqq i \leqq n. \tag{15}$$

Let us define $b_i^j := \max(y_{i-j}, z_j)$, for $i = 1, \cdots, n+j, j = 0, 1, \cdots, m$. We agree that $y_i := 0$ if $i < 1$. Let us call $b^j := \sum_{i=1}^{n+j} b_i^j$ for $j = 0, 1, \cdots, m$. Next, we prove that

$$q_j = b^j - b^{j-1} \quad \text{for } j = 1, \cdots, m. \tag{16}$$

For this, we must compute b_i^j for each $i = 1, \cdots, n+j$ and each $j = 0, 1, \cdots, m$.

If $i \leqq j$, then $b_i^j = \max(y_{i-j}, z_i) = z_i$.

If $i > j$, then there exists an integer k, $1 \leqq k \leqq n$, such that $i = k + j$. Then, by (14),

$$y_k = z_{k+t-1}$$

with $t \in \{1, \cdots, m+1\}$ such that $g_{t-1} + 1 \leqq k \leqq g_t$, and thus

$$b_i^j = \max(y_k, z_{k+j}) = \max(z_{k+t-1}, z_{k+j}).$$

Now, two different cases are possible:

(i) If $j \geqq t-1$, then $b_i^j = z_{k+j} = z_i$,

(ii) If $j < t-1$, then $b_i^j = z_{k+t-1} = y_k = y_{i-j}$.

It is easy to see that (i) holds if and only if $j+1 \leqq i \leqq g_{j+1}+j$ and that (ii) is equivalent to $g_{j+1}+j+1 \leqq i \leqq n+j$. Therefore, $b_i^j = z_i$ for $i = 1, \cdots, g_{j+1}+j$, and $b_i^j = y_{i-j}$ for $i = g_{j+1}+j+1, \cdots, n+j$. Then for $j = 0, 1, \cdots, m$ $b^j = z_1 + \cdots + z_{g_{j+1}+j} + y_{g_{j+1}+1} + \cdots + y_n$, and for $j = 1, \cdots, m$ $b^{j-1} = z_1 + \cdots + z_{g_j+j-1} + y_{g_j+1} + \cdots + y_n = z_1 + \cdots + z_{g_j+j-1} + y_{g_j+1} + \cdots + y_{g_{j+1}} + y_{g_{j+1}+1} + \cdots + y_n = z_1 + \cdots + z_{g_j+j-1} + z_{g_j+j+1} + \cdots + z_{g_{j+1}+j} + y_{g_{j+1}+1} + \cdots + y_n$. Thus $b^j - b^{j-1} = z_{g_j+j} = q_j$ for $j = 1, \cdots, m$ and (16) holds.

Now let us put

$$\psi_i := \phi_0^{z_i}, \qquad 1 \leqq i \leqq n,$$

$$\psi_{n+i} := \phi_0^{z_{n+i}} \lambda^{r_i}, \qquad 1 \leqq i \leqq m.$$

Then for $j = 1, \cdots, m$ we have the following: If $1 \leqq i \leqq n$, then $\beta_i^j = \text{l.c.m.}\,(\alpha_{i-j}, \psi_i) = \phi_0^{\max(y_{i-j}, z_i)}$; if $n+1 \leqq i \leqq n+m$, then $\beta_i^j = \phi_0^{\max(y_{i-j}, z_i)} \lambda^{r_i}$. Thus,

$$\beta_1^j \cdots \beta_{n+j}^j = \phi_0^{\sum_{i=1}^{n} b_i^j} \phi_0^{\sum_{i=n+1}^{n+j} b_i^j} \lambda^{r_1} \cdots \lambda^{r_j},$$

$$\beta_1^j \cdots \beta_{n+j-1}^{j-1} = \phi_0^{\sum_{i=1}^{n} b_i^{j-1}} \phi_0^{\sum_{i=n+1}^{n+j-1} b_i^{j-1}} \lambda^{r_1} \cdots \lambda^{r_{j-1}}.$$

Therefore,

$$\sigma_j = \phi_0^{b^j - b^{j-1}} \lambda^{r_j} = \phi_0^{q_j} \lambda^{r_j},$$

and consequently,

$$d(\sigma_j) = q_j d_0 + r_j = x_j.$$

Moreover, by (15),

$$\psi_i | \alpha_i, \qquad 1 \leqq i \leqq n,$$

and

$$\alpha_i | \phi_0^{z_{i+m}}, \qquad 1 \leqq i \leqq n;$$

as $\psi_{n+i} = \phi_0^{n+i} \lambda^{r_i}$, we have that

$$\alpha_i | \psi_{i+m}, \qquad 1 \leqq i \leqq n.$$

In the general case, there exist irreducible polynomials $\phi_1, \cdots, \phi_p \in \mathbb{F}[\lambda]$ such that

$$\alpha_i = \phi_1^{y_{1i}} \cdots \phi_p^{y_{pi}}, \qquad i = 1, \cdots, n.$$

Let d_1 be the degree of ϕ_1. Let q and r_j the quotient and the remainder of the Euclidean division of x_j by d_1:

$$x_j = q_j d_1 + r_j, \qquad 0 \leqq r_j < d_1 \quad \text{for } j = 1, \cdots, m.$$

In this case it suffices to take

$$\psi_k := \phi_1^{z_{1k}} \cdots \phi_p^{z_{pk}} \quad \text{for } k = 1, \cdots, n,$$

$$\psi_{n+k} := \phi_1^{z_{1,n+k}} \cdots \phi_p^{z_{p,n+k}} \lambda^{r_k} \quad \text{for } k = 1, \cdots, m,$$

where

$$(z_{11}, \cdots, z_{1,n+m}) := (y_{11}, \cdots, y_{1n}) \cup (q_1, \cdots, q_m)$$

$$z_{hk} := y_{h,k-m} \quad \text{for } h = 2, 3, \cdots, p \text{ and } k = 1, \cdots, n+m,$$

agreeing that $y_{hi} := 0$ for $i < 1$.

Actually, from (15) and (16) we deduce

(17) $\quad z_{1i} \leqq y_{1i} \leqq z_{1,i+m} \qquad (i = 1, \cdots, n)$, and

(18) $$x_j = d_1 \left[\sum_{i=1}^{n+j} \max\,(y_{1,i-j}, z_{1i}) - \sum_{i=1}^{n+j-1} \max\,(y_{1,i-j+1}, z_{1i}) \right] + r_j \qquad (j = 1, \cdots, m);$$

and for $h = 2, \cdots, p$ we have that

(19) $\quad z_{hi} \leqq y_{hi} \leqq z_{h,i+m} \qquad (i = 1, \cdots, n)$, and

(20) $\quad \max\,(y_{h,i-j}, z_{hi}) = y_{h,i-j}$ for $i = 1, \cdots, n+j$ and $j = 0, 1, \cdots, m$,

and therefore

$$\sum_{i=1}^{n+j} \max(y_{h,i-j}, z_{hi}) - \sum_{i=1}^{n+j-1} \max(y_{h,i-j+1}, z_{hi}) = 0 \quad \text{for } h = 2, \cdots, p.$$

From (17) and (19) we get (12), and (13) follows from (18) and (20), and the lemma is proved. □

Remark. As by hypothesis $\sum_{i=1}^{n} d(\alpha_i) = n - s$ and $\sum_{j=1}^{m} x_j = s$, from (8) we have that $\sum_{i=1}^{n} d(\gamma_i) = n$.

Proof of Theorem 1. First we will prove that (1) implies (2). In fact, this is an immediate consequence of the definition of feedback equivalence. Actually, for each $F_1 \in \mathbb{F}^{m\times n}$ we have that the pair $(A_1 + B_1F_1, B_1)$ is feedback equivalent to the pair (A_1, B_1). But, on the other hand, this pair is feedback equivalent to (A_2, B_2). Thus there exist matrices $K \in \mathbb{F}^{m\times n}$ and $P \in \mathbb{F}^{n\times n}$, P invertible, such that

$$A_1 + B_1F_1 = P^{-1}(A_2 + B_2K)P.$$

If we take $F_2 := K$ we see that $A_2 + B_2F_2$ is similar to $A_1 + B_1F_1$.

We can prove in the same way that for each $F_2 \in \mathbb{F}^{m\times n}$ there exists $F_1 \in \mathbb{F}^{m\times n}$ such that $A_2 + B_2F_2$ and $A_1 + B_1F_1$ are similar.

Now we will prove that (2) implies (1). To prove that the matrix pairs (A_1, B_1) and (A_2, B_2) are feedback equivalent, it suffices to show that they have the same invariant polynomials and the same controllability indices. Now, two matrices are similar if and only if they have the same determinantal divisors; condition (2) implies that for each $k = 1, \cdots, n$ the sets of polynomials $\{D_k(A_1 + B_1F_1) | F_1 \in \mathbb{F}^{m\times n}\}$ and $\{D_k(A_2 + B_2F_2) | F_2 \in \mathbb{F}^{m\times n}\}$ are equal. Hence, applying Theorem 3, we have that the invariant polynomials of the pairs (A_1, B_1) and (A_2, B_2) are the same.

Let $(k_1^1, \cdots, k_m^1)$ with $k_1^1 \geqq \cdots \geqq k_m^1 \geqq 0$ be the controllability indices of (A_1, B_1), and $(k_1^2, \cdots, k_m^2)$ with $k_1^2 \geqq \cdots \geqq k_m^2 \geqq 0$ those of (A_2, B_2). Since the sum of the controllability indices and the degrees of the invariant polynomials of a pair is the number n of its rows, we have that

$$\sum_{i=1}^{m} k_i^1 = s = \sum_{i=1}^{m} k_i^2.$$

Let us consider any m-tuple of integers $(x_m, \cdots, x_1)$, $x_m \geqq \cdots \geqq x_1 \geqq 0$, such that

$$(k_1^1, \cdots, k_m^1) \prec (x_m, \cdots, x_1). \tag{21}$$

By Lemma 4, for this m-tuple $(x_m, \cdots, x_1)$ there exist n monic polynomials $\gamma_1 | \cdots | \gamma_n$ of $\mathbb{F}[\lambda]$ satisfying (7) and (8). Then applying Theorem 2 to the pair (A_1, B_1) there exists a matrix $F_1 \in \mathbb{F}^{m\times n}$ such that $A_1 + B_1F_1$ have $\gamma_1, \cdots, \gamma_n$ as invariant polynomials. By (2), for this matrix F_1 there exists a matrix $F_2 \in \mathbb{F}^{m\times n}$ such that $A_1 + B_1F_1$ and $A_2 + B_2F_2$ are similar. Thus, $\gamma_1, \cdots, \gamma_n$ are also the invariant polynomials of $A_2 + B_2F_2$. Applying again Theorem 2 it follows that $\gamma_1, \cdots, \gamma_n$ necessarily satisfy

$$(k_1^2, \cdots, k_m^2) \prec (d(\sigma_m), \cdots, d(\sigma_1)).$$

By (8), we have that

$$(k_1^2, \cdots, k_m^2) \prec (x_m, \cdots, x_1). \tag{22}$$

In an analogous way we prove that (2) implies (21). Therefore, by Proposition 5, we have that

$$(k_1^1, \cdots, k_m^1) = (k_1^2, \cdots, k_m^2).$$

Acknowledgments. We thank the referees, whose comments have allowed us to improve the presentation of this paper.

REFERENCES

[1] S. BARNETT, *Polynomials and Linear Control Systems*, Marcel Dekker, New York, 1983.
[2] S. BARNETT AND R. G. CAMERON, *Introduction to Mathematical Control Theory*, 2nd ed., Clarendon Press, Oxford, 1985.
[3] P. BRUNOVSKY, *A classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173-188.
[4] J. DIEUDONNE, *Calcul Infinitésimal*, Hermann, Paris, 1968.
[5] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Invariant Subspaces of Matrices with Applications*, John Wiley, New York, 1986.
[6] R. E. KALMAN, *Kronecker invariants and feedback*, in Ordinary Differential Equations, L. Weiss, ed., Academic Press, New York, 1972, pp. 459-471.
[7] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Thomas Nelson and Sons Ltd., London, 1970.
[8] A. TANNENBAUM, *Invariance and System Theory: Algebraic and Geometric Aspects*, Lecture Notes in Mathematics 845, Springer-Verlag, Berlin, Heidelberg, New York, 1981.
[9] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer-Verlag, New York, 1985.
[10] I. ZABALLA, *Matrices with prescribed rows and invariant factors*, Linear Algebra Appl., 87 (1987), pp. 113-146.
[11] ———, *Interlacing and majorization in invariant factors assignment problems*, Linear Algebra Appl., 121 (1989), pp. 409-421.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 1113-1135, September 1990

005

AN EXTENSION OF THE MAXIMUM PRINCIPLE FOR A CLASS OF OPTIMAL CONTROL PROBLEMS IN INFINITE-DIMENSIONAL SPACES*

N. BASILE† AND M. MININNI‡

Abstract. An extension of the Pontryagin maximum principle and of the transversality conditions for a class of optimal control problems for a system of a parabolic equation and an ordinary differential equation in a Hilbert space are given. In particular, the time optimal problem for some of these systems is considered. As an application the optimal control of the diffusion of a class of epidemics is studied.

Key words. optimal control problems, maximum principle, transversality conditions, parabolic equations

AMS(MOS) subject classification. 49B22

1. Introduction. It is well known (see [3]) that the Pontryagin maximum principle can be extended to optimal control problems for evolution equations in infinite-dimensional vector spaces with free final state. However, if there is a constraint on the final state, then the maximum principle does not hold in general (see [2, p. 251]).

Recently, Fattorini and Frankowska [12], [13], by making use of the Ekeland's variational principle, give some conditions on the reachable set and on the target set in order to get an extension of the maximum principle to a large class of problems in infinite-dimensional spaces.

On the other hand, Li and Yao [16] by making use of the Eidelheit separation theorem and of an extension of the Uhl's theorem extend the maximum principle and the transversality condition for integral systems with time lags, when the target set is convex and the final time T is fixed. In the same frame is a recent paper by Li and Chow [15] for optimal periodic control of functional differential equations.

In this paper, by using the same arguments as Li and Yao, we extend the maximum principle and the transversality conditions to optimal control problems for a system of a parabolic equation and an ordinary differential equation in a Hilbert space. Moreover, for a class of such problems we can consider the case when the final time T is unknown; in particular, we can consider the time optimal problem for some of these systems.

As an application we study the optimal control of the diffusion of a class of epidemics that initially motivated this research.

The paper is organized as follows. In § 2 we introduce the notation and state the main results, in § 3 we collect some remarks and prove some useful lemmata, and in § 4 we prove a lemma that has a crucial role in the proof of the theorems stated in § 2. Finally, in § 5 we prove these theorems, and in § 6 we give the mentioned application to the study of epidemics. As an appendix to the paper we prove in § 7 an existence and uniqueness result for linear evolution equations, which is used to solve the adjoint equations. We suspect that the result is known, but we have not been able to find it explicitly.

2. Notation and statement of the main results. Let X_1, H be real separable Hilbert spaces such that $H \subset X_1 \simeq X_1' \subset H'$, H dense in X_1, endowed, respectively, with the

* Received by the editors December 2, 1988; accepted for publication (in revised form) October 30, 1989. This work was supported by M.P.I. of Italy, "Fondi 40%: Equazioni differenziali e calcolo delle variazioni" and "Fondi 60%: Università di Bari e Università della Calabria."

† Dipartimento di Matematica dell'Università, via Guistino Fortunato, 70125, Bari, Italy.

‡ Dipartimento di Matematica dell'Università della Calabria, 87036 Arcavacata di Rende (Cosenza), Italy.

inner products $(\cdot,\cdot)$, $((\cdot,\cdot))$ and the norms $|\cdot|$ and $\|\cdot\|$. Moreover, let us denote by $\langle\cdot,\cdot\rangle$ the canonical pairing between H and H', and obviously we have $\langle x, y\rangle = (x, y)$ for all $x \in X_1$, $y \in H$. Let $W(0, T; H)$ be the Hilbert space of the (classes of) functions $x \in L^2(0, T; H)$ whose derivatives x' in the sense of distributions belong to $L^2(0, T; H')$, endowed with the norm

$$\|x\| = \left(\int_0^T [\|x(t)\|_H^2 + \|x'(t)\|_{H'}^2]\, dt \right)^{1/2}.$$

Moreover, let X_2 be another real separable Hilbert space whose norm is still denoted by $|\cdot|$, and let us put $X = X_1 \times X_2$ endowed with the canonical norm $|x| = (|x_1|^2 + |x_2|^2)^{1/2}$. Finally, let U be a subset of a Banach space Z, let $\mathcal{U}(0, T)$ be the space of the strongly measurable functions from $[0, T]$ to Z such that $u(t) \in U$ for almost all t, and let $\mathcal{U}_{ad}(0, T)$ be a subset of $\mathcal{U}(0, T)$ such that:

($\mathcal{U}$) If $v_1, v_2, \cdots, v_k \in \mathcal{U}_{ad}(0, T)$ and $E_1, E_2, \cdots, E_k$ is a measurable partition of $[0, T]$, then $\sum_{i=1}^k \chi_i v_i \in \mathcal{U}_{ad}(0, T)$ (where χ_i is the characteristic function of E_i).

Now consider a linear continuous self-adjoint operator A from H to H' such that for some $\alpha > 0$, $\beta \in \mathbf{R}$ we have

$$\langle Ax, x\rangle \geqq \alpha \|x\|^2 - \beta |x|^2 \quad \text{for all } x \in H, \tag{2.1}$$

let Φ be a Frechét differentiable mapping from $\mathbf{R}_+ \times X$ into $\mathbf{R}$, and let $f = (f_1, f_2) = f(t, x, u)$ be a mapping from $[0, +\infty[\times X_1 \times X_2 \times U$ to $H' \times X_2$ satisfying the following conditions for all $T > 0$:

(f.1) f is Frechét differentiable with respect to $x = (x_1, x_2)$;

(f.2) for all $(x, u) \in C(0, T; X) \times \mathcal{U}_{ad}(0, T)$ we have for some $q > 2$:
- (1) $f(\cdot, x(\cdot), u(\cdot)) \in L^q(0, T; H') \times L^1(0, T; X_2)$,
- (2) $f_x(\cdot, x(\cdot), u(\cdot))(z(\cdot))$ is strongly measurable for all $z \in L^\infty(0, T; X)$,
- (3) $\|f_x(t, (x(t), u(t)))\| \leqq M(t)$ almost everywhere in $[0, T]$ for some $M \in L^q(0, T; \mathbf{R})$;

(f.3) for all $\xi \in X$ there exist $r > 0$, $\varphi : [0, r] \to \mathbf{R}_+$ and $L : [0, T] \times U \to \mathbf{R}_+$ such that:
- (1) for all $u \in \mathcal{U}_{ad}(0, T)$ the mapping $L(\cdot, u(\cdot))$ belongs to $L^2(0, T; \mathbf{R})$,
- (2) $\varphi(r) = o(r)$ as $r \to 0$,
- (3) for all $t \in [0, T]$, $u \in U$, x, y in the ball with center ξ and radius r we have $\|f(t, x, u) - f(t, y, u)\| \leqq L(t, u)|x - y|$, and $\|(f_x(t, x, u) - f_x(t, \xi, u))(x - \xi)\| \leqq L(t, u)\varphi(|x - \xi|)$.

We are interested in the study of the following optimal control problem:

(P) Minimize the functional $J(T, x, u) = \Phi(T, x(T))$ with (x, u), $x = (x_1, x_2)$ such that

$$\begin{aligned}
&x_1 \in W(0, T; H) \cap C(0, T; X_1), \quad x_2 \in AC(0, T; X_2), \quad u \in \mathcal{U}_{ad}(0, T),\\
&x_1'(t) + Ax_1(t) = f_1(t, x(t), u(t)) \quad \text{a.e. in } [0, T],\\
&x_2'(t) = f_2(t, x(t), u(t)) \quad \text{a.e. in } [0, T],\\
&x_1(0) = x_1^0, \qquad x_2(0) = x_2^0,
\end{aligned} \tag{2.2}$$

satisfying the constraint on the final state

$$(T, x(T)) \in \mathcal{B} \subset \mathbf{R}_+ \times X. \tag{2.3}$$

To this end let us consider the Hamiltonian function

$$H(t, x, u, p) = \langle f(t, x, u), p\rangle = \langle f_1(t, x_1, x_2, u), p_1\rangle + (f_2(t, x_1, x_2, u), p_2)$$

for all $t \in [0, T]$, $x = (x_1, x_2) \in X$, $p = (p_1, p_2) \in H \times X_2$, $u \in U$. Then we have the following theorems.

THEOREM 2.1. *Let (T, x, u) be an optimal solution of* (P) *in the case when the final time T is fixed, (i.e., $\mathcal{B} = \{T\} \times B$ with $B \subset X$), and assume that B is convex with nonempty interior. Then there exist $\lambda_0 \in \{0, 1\}$, $p_1 \in (W(0, T; H) + W^{1,q^*}(0, T; X_1)) \cap C(0, T; X_1) \cap L^2(0, T; H)$, $q^* = 2q/(2+q)$, and $p_2 \in AC(0, T; X_2)$, which satisfy the nondegeneracy condition*

$$(\lambda_0, p(T)) \neq (0, 0), \tag{2.4}$$

the adjoint equations

$$\begin{aligned} &p_1'(t) - Ap_1(t) = -H_{x_1}(t, x(t), u(t), p(t)) = -B_{11}^*(t)p_1(t) - B_{21}^*(t)p_2(t) \quad \text{in } [0, T], \\ &p_2'(t) = -H_{x_2}(t, x(t), u(t), p(t)) = -B_{12}^*(t)p_1(t) - B_{22}^*(t)p_2(t) \quad \text{in } [0, T] \end{aligned} \tag{2.5}$$

(where $B_{ij}(t) = (f_i)_{x_j}(t, x(t), u(t))$, $t \in [0, T]$, $i, j \in \{1, 2\}$ and $*$ denotes the adjoint), the maximum principle

$$\int_0^T H(t, x(t), u(t), p(t))\, dt \leqq \int_0^T H(t, x(t), v(t), p(t))\, dt \quad \text{for all } v \in \mathcal{U}_{\text{ad}}(0, T), \tag{2.6}$$

and the transversality condition

$$(p(T) - \lambda_0 \Phi_x(T, x(T)), \xi - x(T)) \leqq 0 \quad \text{for all } \xi \in B. \tag{2.7}$$

Moreover, we have $p_1 \in W(0, T; H)$ if (f.2)(3) *holds with $q = +\infty$ and $p_1 \in C(0, T; H) \cap W^{1,2}(0, T; X_1)$ if $p_1(T) \in H$.*

THEOREM 2.2. *Assume now that (T, x, u) is an optimal solution of* (P) *in the free final time case, and assume that*:

(1) *Φ does not depend on x_1,*

(2) *the target set $\mathcal{B}$ has the form $\mathcal{B} = \mathbf{R}_+ \times X_1 \times B_2$ and B_2 is a convex subset of X_2 with nonempty interior,*

(3) *there exists $\lim_{t \to T^-} f_2(t, x(t), u(t)) = f_2^-(T, x(T), u(T))$.*

Then there exist $\lambda_0 \in \{0, 1\}$, $p = (p_1, p_2) \in (C(0, T; H) \cap W^{1,2}(0, T; X_1)) \times AC(0, T; X_2)$ satisfying (2.4), (2.5), (2.6), *and the transversality conditions*

$$p_1(T) = 0, \quad (p_2(T) - \lambda_0 \Phi_{x_2}(T, x_2(T)), \xi_2 - x_2(T)) \leqq 0 \quad \text{for all } \xi_2 \in B_2, \tag{2.7a}$$

$$H(T, x(T), u(T), p(T)) = (f_2^-(T, x(T), u(T)), p(T)) \leqq -\lambda_0 \Phi_t(T, x(T)). \tag{2.7b}$$

THEOREM 2.3. *If $\Phi(t, x) = t$, (i.e., if (T, x, u) is a solution for a time optimal control problem for* (2.2)-(2.3)), *then the assertion of Theorem* 2.3 *holds with condition* (2.7b) *replaced by*

$$H(T, x(T), u(T), p(T)) = (f_2^-(T, x(T), u(T)), p(T)) = -\lambda_0. \tag{2.7b'}$$

An interesting variant of problem (P) is the following:

(P') Minimize the functional $J(T, x, u) = \Phi(T, x(T))$, where (x, u) is a solution of (2.2) satisfying the constraint on the final state (2.3) and a finite number of isoperimetric constraints of the form

$$\int_0^T h_i(u(t))\, dt \leqq \text{const.} = c_i, \qquad i = 1, \cdots, l \tag{2.8}$$

where $h_1, h_2, \cdots, h_l : Z \to \mathbf{R}_+ \cup \{+\infty\}$ and $\mathcal{U}_{ad}(0, T)$ is the set of all $u \in \mathcal{U}(0, T)$ such that $h_j(u(\cdot))$ is summable for all j. (Note that $\mathcal{U}_{ad}(0, T)$ so defined satisfies condition $(\mathcal{U})$.)

Then we have the following corollaries.

COROLLARY 2.4. *Under the assumptions of Theorem* 2.2, *let* (T, x, u) *be a solution of* (P'). *Then there exist* $\lambda_0 \in \{0, 1\}$, $\mu_1, \mu_2, \cdots, \mu_l \in \mathbf{R}$ *and* $p = (p_1, p_2)$, *which satisfy the nondegeneracy condition*

$$(\lambda_0, \mu_1, \cdots, \mu_l, p(T)) \neq 0, \tag{2.4'}$$

the adjoint equations (2.5), *the maximum principle*

$$\int_0^T \left(H(t, x(t), u(t), p(t)) + \sum_{i=1}^{l} \mu_i h_i(u(t)) \right) dt \leqq \int_0^T \left(H(t, x(t), v(t), p(t)) + \sum_{i=1}^{l} \mu_i h_i(v(t)) \right) dt \quad \text{for all } v \in \mathcal{U}_{ad}(0, T), \tag{2.6'}$$

and the transversality conditions (2.7a), (2.7b), *and*

$$\mu_i \geqq 0 \quad \text{and} \quad \mu_i = 0 \quad \text{if} \int_0^T h_i(u(t))\, dt < c_i. \tag{2.7c}$$

COROLLARY 2.5. *The assertion of Corollary* 2.4 *holds with* (2.7b) *replaced by* (2.7b') *in the case of a time optimal problem, i.e., when* $\Phi(t, x) = t$.

Remark 2.6. The maximum principle (2.6) takes the more familiar form:

$$H(t, x(t), u(t), p(t)) = \min_{v \in U} H(t, x(t), v, p(t)) \quad \text{a.e. in } [0, T], \tag{2.9}$$

if we assume that $\mathcal{U}_{ad}(0, T)$ contains the constant functions $v(t) \equiv v$ for all $v \in U$ and

(2.10) there exists a subset S of $[0, T]$ such that meas $S = 0$ and for all $v \in U$, $t \notin S$ we have $\lim_{\rho \to 0} 1/\rho \int_{t-\rho}^{t} \langle f(s, x(s), v), p(s) \rangle\, ds = \langle f(t, x(t), v), p(t) \rangle$.

(Note that (2.10) holds, for example, if the mapping $f(\cdot, x(\cdot), v)$ is continuous for all v; in particular, it holds in the case where f is independent of t.)

In fact, let us fix $t \in]0, T[\setminus S$ such that t is a Lebesgue point of $\langle f(\cdot, x(\cdot), u(\cdot)), p(\cdot) \rangle$ and $v \in U$, and for $\rho > 0$ sufficiently small let us put

$$v_\rho(s) = \begin{cases} u(s) & \text{for } s \notin [t-\rho, t], \\ v & \text{for } s \in [t-\rho, t]. \end{cases}$$

Then we have that $v_\rho \in \mathcal{U}_{ad}(0, T)$, and therefore

$$0 \leqq \frac{1}{\rho} \int_0^T (H(s, x(s), v_\rho(s), p(s)) - H(s, x(s), u(s), p(s)))\, ds$$

$$= \frac{1}{\rho} \int_{t-\rho}^{t} \langle f(s, x(s), v) - f(s, x(s), u(s)), p(s) \rangle\, ds;$$

hence by passing to the limit as $\rho \to 0^+$, we have $0 \leqq \langle f(t, x(t), v) - f(t, x(t), u(t)), p(t) \rangle$, i.e., (2.9).

In a similar way we prove that (2.6') takes the more familiar form

$$H(t, x(t), u(t), p(t)) + \sum_{i=1}^{l} \mu_i h_i(u(t)) = \min_{v \in U} H(t, x(t), v, p(t)) + \sum_{i=1}^{l} \mu_i h_i(v), \tag{2.9'}$$

if (2.10) holds. (Note that $v(t) \equiv v \in \mathcal{U}_{ad}(0, T)$ if and only if $h_i(v) < +\infty$ for all i, but we have $H(t, x(t), v, p(t)) + \sum_{i=1}^{l} \mu_i h_i(v) = +\infty > H(t, x(t), u(t), p(t)) + \sum_{i=1}^{l} \mu_i h_i(u(t))$ if $h_i(v) = +\infty$ for some i.)

Remark 2.7. With obvious modifications in the proof we can obtain the preceding results also in the case when the operator A is replaced by a family $(A(t))_{t\in[0,T]}$ of linear bounded self-adjoint operators from H to H' that are uniformly coercive with respect to t and such that $\|A(t) - A(s)\| \leqq c|t - s|^{\alpha}$ for some $c > 0$, $\alpha \in]0, 1[$.

This can be used, for example, in the applications given in § 6, where we can reasonably assume that the diffusion of the epidemic is influenced by seasonal factors.

Remark 2.8. Note that (f.2) holds, for example, if

(1) $f(t, x, u)$ and $f_x(t, x, u)y$ are strongly measurable in t for all (x, u, y) and continuous in (x, u, y) for almost all t;

(2) for all compact subsets K of X there exist $q > 2$ and $M \in L^1(0, T)$ such that

$$\|f_1(t, x, u)\|^q + |f_2(t, x, u)| + \|f_x(t, x, u)\|^q \leqq M(t)$$

for all $x \in K$, $u \in U$ and for almost all $t \in [0, T]$.

3. Some generalities on the evolution equations and preliminary lemmata. First of all note that without loss of generality we can assume $\beta = 0$ in (2.1).

Remark 3.1. It is well known (see [17, p. 116]) that $W(0, T; H)$ is contained in $C(0, T; X_1)$ in the sense that if $x \in W(0, T; H)$, then there exists $\tilde{x} \in C(0, T; X_1)$ such that $\tilde{x}(t) = x(t)$ almost everywhere. Moreover (see [17, pp. 116–124]), for every $g \in L^2(0, T; H')$, $x_0 \in X_1$, the evolution Cauchy problem

$$\begin{aligned} &x'(t) + Ax(t) = g(t), \qquad t \in [0, T], \\ &x(0) = x_0 \end{aligned} \tag{3.1}$$

has a unique solution $x \in W(0, T; H) \cap C(0, T; X_1)$ and (for some $c > 0$) we have

$$\|x\| \leqq c(|x_0| + \|g\|). \tag{3.2}$$

Finally (see [18, p. 76]), $-A$ is the infinitesimal generator of an analytic semigroup G on X_1 and H' such that for some $c > 0$ we have

$$|G(t)x_0| \leqq |x_0|, \quad \|G(t)x_0\|_{H'} \leqq c\|x_0\|_{H'}, \quad |G(t)x_0| \leqq ct^{-1/2}\|x_0\|_{H'}; \tag{3.3}$$

moreover, the unique solution $x = x(t)$ of (3.1) can be represented in the form

$$x(t) = G(t)x_0 + \int_0^t G(t - s)g(s)\, ds, \qquad t \in [0, T]. \tag{3.4}$$

From (3.4), (3.2) and the fact that

$$|x(t)|^2 - |x(0)|^2 = \int_0^t (|x(s)|^2)'\, ds \leqq \int_0^t [\|x(s)\|_H^2 + \|x'(s)\|_{H'}^2]\, dt,$$

it follows that

$$\left| \int_0^t G(t - s)g(s)\, ds \right|^2 \leqq c^2 \int_0^t \|g(s)\|_{H'}^2\, ds \quad \text{for all } t \in [0, T], \quad g \in L^2(0, T; H'). \tag{3.5}$$

Remark 3.2. For all $t \in [0, T]$ let $B(t)$ be a linear continuous mapping from X_1 in H' such that $B(t)x$ is strongly measurable in t for all x and $\|B(t)\| \leqq L(t)$ for some square summable function L. Then by means of the usual successive approximation technique we easily show that for every $\xi \in X_1$, $g \in L^2(0, T; H')$ the Cauchy problem

$$\begin{aligned} &y'(t)+Ay(t)=B(t)y(t)+g(t), \qquad t\in[0,T],\\ &y(0)=\xi, \end{aligned} \tag{3.6}$$

has a unique solution in $W(0, T; H)\cap C(0, T; X_1)$.

Remark 3.3. Under the assumptions of the preceding remark, if $L\in L^q(0, T)$, $q>2$, then we can prove (see Theorem 7.1 in the Appendix) that the linear Cauchy problem

$$\begin{aligned} &p'(t)+Ap(t)=B^*(t)p(t), \qquad t\in[0,T],\\ &p(0)=\eta \end{aligned}$$

(where $B^*(t): H\to X_1$ denotes the adjoint operator of $B(t)$), has a unique solution $p\in C(0, T; X_1)\cap L^2(0, T; H)$ with $p-G(\cdot)\eta\in W^{1,q^*}(0, T; X_1)$, $q^*=2q/(2+q)$.

Actually, $p-G(\cdot)\eta\in W^{1,2}(0, T; X_1)\cap C(0, T; H)$ (and therefore $p\in W(0, T; H)$) if $q=+\infty$. Moreover, if $\eta\in H$ then we have $p\in W^{1,2}(0, T; X_1)\cap C(0, T; H)$, also when $q=2$.

Finally note that by means of the change of variables $s=T-t$ we also have that the problem

$$\begin{aligned} &p'(t)-Ap(t)=-B^*(t)p(t), \qquad t\in[0,T],\\ &p(T)=\eta \end{aligned} \tag{3.7}$$

has a unique solution as described above.

Remark 3.4. Under the assumptions of Remarks 3.2 and 3.3, if y, p are solutions of (3.6) and (3.7), respectively, then we have that

$$(p(T), y(T))-(p(0), y(0))=\int_0^T \langle g(s), p(s)\rangle\, ds.$$

This easily follows from the fact that the mapping $\gamma(t)=(p(t), y(t))$ is absolutely continuous and (almost everywhere in $[0, T]$) we have

$$\gamma'(t)=\langle y'(t), p(t)\rangle+\langle p_0'(t), y(t)\rangle+(p_1'(t), y(t)),$$

where $p_0\in W(0, T; H)$, $p_1\in L^2(0, T; H)\cap W^{1,q^*}(0, T; X_1)$, are solutions of the Cauchy problems $p_0'-Ap_0=0$, $p_0(T)=p(T)$, and $p_1'-Ap_1=-B^*(\cdot)p$, $p_1(T)=0$.

Remark 3.5. With the same arguments used in the proof of Lemma 6.2 of [14, p. 36], we prove the following Gronwall type inequality.

Assume that $0\leqq\varphi(t)\leqq M+\int_0^t L(s)\varphi(s)\, ds$ for all $t\in[0, T]$ where $M>0$, $L\in L^1(0, T; \mathbf{R})$, $L\geqq 0$, and $\varphi\in L^\infty(0, T; \mathbf{R})$. Then we have $\varphi(t)\leqq M\exp(\int_0^t L(s)\, ds)$ for all $t\in[0, T]$.

The following lemma whose proof can be found in [15] has a crucial role in the following.

LEMMA 3.6. *Let $\lambda_1\geqq 0, \cdots, \lambda_k\geqq 0$ with $\sum_{i=1}^k \lambda_i=1$, let X be a Banach space, and let $g_1, \cdots, g_k\in L^1(0, T; X)$. Then for all $\varepsilon\in\,]0, 1]$ there exists a family of measurable mutually disjoint subsets $E_1, \cdots, E_k$ of $[0, T]$ such that $\sum_{i=1}^k \operatorname{meas}(E_i)=\varepsilon T$ and*

$$\left|\varepsilon\int_0^t \sum_{i=1}^k \lambda_i g_i(s)\, ds-\sum_{i=1}^k\int_{E_i\cap[0,t]} g_i(s)\, ds\right|<\varepsilon^2 \quad \text{for all } t\in[0,T].$$

LEMMA 3.7. *Let $\lambda_1\geqq 0, \cdots, \lambda_k\geqq 0$ with $\sum_{i=1}^k \lambda_i=1$, let X be a Banach space, and let $g_1, \cdots, g_k$ be mappings from $[0, T]\times[0, T]$ to X such that*

(1) $g_i(t, \cdot)\in L^1(0, T; X)$

(2) $\lim_{t\to t_0}\int_0^T |g_i(t, s)-g_i(t_0, s)|\, ds=0$ *for all* $t_0\in[0, T]$.

Then for all $\varepsilon \in]0, 1]$ *there exists a family of measurable mutually disjoint subsets* $E_1, \cdots, E_k$ *of* $[0, T]$ *such that* $\sum_{i=1}^{k} \text{meas}\,(E_i) = \varepsilon T$ *and*

$$\left| \varepsilon \int_0^t \sum_{i=1}^{k} \lambda_i g_i(t, s)\, ds - \sum_{i=1}^{k} \int_{E_i \cap [0,t]} g_i(t, s)\, ds \right| < 2\varepsilon^2 \quad \textit{for all } t \in [0, T].$$

Proof. In fact, by assumption the mappings $t \to g_i(t, \cdot)$ are continuous (and therefore uniformly continuous) from $[0, T]$ in $L^1(0, T; X)$. Hence for $\varepsilon > 0$ fixed there exist $0 = t_0 < t_1 < \cdots < t_l = T$ such that

$$\int_0^T |g_i(t', s) - g_i(t'', s)|\, ds < \frac{\varepsilon^2}{2k} \quad \text{for all } t', t'' \in [t_{j-1}, t_j], \quad j = 1, \cdots, l. \tag{3.8}$$

Now let us put $h_i(s) = (g_i(t_0, s), \cdots, g_i(t_l, s))$ for all $i = 1, \cdots, k$, $s \in [0, T]$. Then it is evident that $h_i \in L^1(0, T; X^{l+1})$. Therefore by Lemma 3.6 there exists a family of measurable mutually disjoint subsets $E_1, \cdots, E_k$ of $[0, T]$ such that $\sum_{i=1}^{k} \text{meas}\,(E_i) = \varepsilon T$ and $|\varepsilon \int_0^t \sum_{i=1}^{k} \lambda_i h_i(s)\, ds - \sum_{i=1}^{k} \int_{E_i \cap [0,t]} h_i(s)\, ds| < \varepsilon^2$, i.e.,

$$\left| \varepsilon \int_0^t \sum_{i=1}^{k} \lambda_i g_i(t_j, s)\, ds - \sum_{i=1}^{k} \int_{E_i \cap [0,t]} g_i(t_j, s)\, ds \right| < \varepsilon^2 \quad \text{for all } t, j.$$

Then for all $t \in [t_{j-1}, t_j], j = 1, \cdots, l$ we have

$$\begin{aligned} &\left| \varepsilon \int_0^t \sum_{i=1}^{k} \lambda_i g_i(t, s)\, ds - \sum_{i=1}^{k} \int_{E_i \cap [0,t]} g_i(t, s)\, ds \right| \\ &\quad \leqq \varepsilon \int_0^t \sum_{i=1}^{k} \lambda_i |g_i(t, s) - g_i(t_j, s)|\, ds + \sum_{i=1}^{k} \int_{E_i \cap [0,t]} |g_i(t_j, s) - g_i(t, s)|\, ds \\ &\qquad + \left| \varepsilon \int_0^t \sum_{i=1}^{k} \lambda_i g_i(t_j, s)\, ds - \sum_{i=1}^{k} \int_{E_i \cap [0,t]} g_i(t_j, s)\, ds \right| < 2\varepsilon^2. \end{aligned}$$

LEMMA 3.8. *Let* G *be the semigroup generated by* $-A$ (*see Remark* 3.1), *let* $x \in L^q(0, T; H')$, $q > 2$, *and let us put*

$$y(t, s) = \begin{cases} G(t-s)x(s) & \textit{for } t \geqq s, \\ 0 & \textit{for } t < s. \end{cases}$$

Then $y \in L^1([0, T] \times [0, T], X_1)$ *and for almost all* $t_0 \in [0, T]$ *we have*

$$\lim_{t \to t_0} \int_0^T |y(t, s) - y(t_0, s)|\, ds = 0.$$

Proof. The first part of the assertion follows from the fact that evidently y is strongly measurable and we have by (3.3)

$$\begin{aligned} \int_0^T \int_0^T |y(t, s)|\, ds\, dt &= \int_0^T \int_0^t |G(t-s)x(s)|\, ds\, dt \\ &\leqq c \int_0^T \int_0^t (t-s)^{-1/2} \|x(s)\|_{H'}\, ds\, dt \\ &\leqq Tc' \left(\int_0^T \|x(s)\|_{H'}^q\, ds \right)^{1/q} < +\infty, \end{aligned}$$

where $c' = c(\int_0^T \sigma^{-q'/2}\, d\sigma)^{1/q'}$ and $(1/q) + (1/q') = 1$.

Now, to prove the second part of the assertion, let us fix t_0. Then for $t > t_0$ we have

$$\int_0^T |y(t, s) - y(t_0, s)|\, ds = \int_0^{t_0} |[G(t-s) - G(t_0 - s)]x(s)|\, ds + \int_{t_0}^t |G(t-s)x(s)|\, ds.$$

The first addend tends to zero as $t \to t_0$ by the Lebesgue convergence theorem, since by (3.3) we have

$$|[G(t-s)-G(t_0-s)]x(s)| = |(G(t-t_0)-I)(G(t_0-s)x(s))| \leqq 2|G(t_0-s)x(s)|,$$

whereas the second addend tends to zero since by (3.5) we have

$$\int_{t_0}^{t} |G(t-s)x(s)|\,ds \leqq c\int_{t_0}^{t} (t-s)^{-1/2}\|x(s)\|_{H'}\,ds \leqq c\|x\|_{L^q}\left(\int_0^{t-t_0} \sigma^{-q'/2}\,d\sigma\right)^{1/q'}.$$

On the other hand, for $t < t_0$, we have

$$\int_0^T |y(t,s)-y(t_0,s)|\,ds = \int_0^t |G(t-s)[I-G(t_0-t)]x(s)|\,ds + \int_t^{t_0} |G(t_0-s)x(s)|\,ds.$$

Once again the second addend evidently tends to zero, whereas the first one tends to zero by (3.3), the Hölder inequality and the Lebesgue convergence theorem. (Actually it is majorized by $c'(\int_0^T \|[I-G(t_0-t)]x(s)\|_{H'}^q\,ds)^{1/q}$ and $\|[I-G(t_0-t)]\cdot x(s)\|_{H'} \leqq (1+c)\|x(s)\|_{H'}$ by (3.3).)

4. A basic lemma. From now on let us fix a solution (x, u) of (2.2) and let us denote by $\mathbf{A}$, $\mathbf{G}(t)$, and $\mathbf{B}(t)$, respectively, the matrices

$$\begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}, \quad \begin{pmatrix} G(t) & 0 \\ 0 & I \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} B_{11}(t) & B_{12}(t) \\ B_{21}(t) & B_{22}(t) \end{pmatrix},$$

where $B_{ij}(t) = (f_i)_{x_j}(t, x(t), u(t))$, for all $t \in [0, T]$, $i, j \in \{1, 2\}$). Then by (f.2)(3) and Remark 3.2 with A, B, ξ, g replaced by $\mathbf{A}$, $\mathbf{B}$, $\xi = (\xi_1, \xi_2)$, $g = (g_1, g_2)$, we have that for all $v \in \mathcal{U}_{\text{ad}}(0, T)$ the linear problem

$$\begin{aligned} &y_1'(t) + Ay_1(t) = B_{11}(t)y_1(t) + B_{12}(t)y_2(t) + f_1(t, x(t), v(t)) - f_1(t, x(t), u(t)), \\ &y_2'(t) = B_{21}(t)y_1(t) + B_{22}(t)y_2(t) + f_2(t, x(t), v(t)) - f_2(t, x(t), u(t)), \\ &y_1(0) = 0, \qquad y_2(0) = 0, \end{aligned} \tag{4.1}$$

has a unique solution $y^v = (y_1^v, y_2^v)$. Then we have the following lemma.

LEMMA 4.1. *Let us fix $v_1, \cdots, v_k \in \mathcal{U}_{\text{ad}}(0, T)$ and $\lambda_1 \geqq 0, \cdots, \lambda_k \geqq 0$ with $\sum_{i=1}^k \lambda_i = 1$, and for all i let us denote by y^i the unique solution of* (4.1) *with $v = v_i$. Then there exists $\varepsilon_0 > 0$ such that for all $\varepsilon \in]0, \varepsilon_0[$ there exists $(x_\varepsilon, u_\varepsilon)$ solution of* (2.2) *such that*

$$z_\varepsilon = \frac{x_\varepsilon - x}{\varepsilon} \text{ converges to } y = \sum_{i=1}^k \lambda_i y^i \text{ in } C(0, T; X_1 \times X_2) \text{ as } \varepsilon \to 0.$$

Proof. First of all note that for all $i = 1, \cdots, k$ we have

$$y^i(t) = \int_0^t \mathbf{G}(t-s)[\mathbf{B}(s)y^i(s) + f(s, x(s), v_i(s)) - f(s, x(s), u(s))]\,ds. \tag{4.2}$$

Now for all $i = 1, \cdots, k$ let us put

$$h_i(t, s) = \begin{cases} \mathbf{G}(t-s)[f(s, x(s), v_i(s)) - f(s, x(s), u(s))] & \text{if } s \leqq t, \\ 0 & \text{if } s > t. \end{cases}$$

By (f.2)(1) and Lemma 3.8, we have that

$$h_i \in L^1([0, T] \times [0, T], X_1 \times X_2) \quad \text{and} \quad \lim_{t \to t_0} \int_0^T |h_i(t, s) - h_i(t_0, s)|\,ds = 0.$$

Hence by Lemma 3.7 for all $\varepsilon > 0$ there exists a family of mutually disjoint measurable subsets $E_1, \cdots, E_k$ of $[0, T]$ such that $\sum_{i=1}^{k} \text{meas}(E_i) = \varepsilon T$ and (for all $t \in [0, T]$)

$$(4.3)\quad \sum_{i=1}^{k} \int_{E_i \cap [0,t]} h_i(t, s)\, ds = \varepsilon \left(\sum_{i=1}^{k} \lambda_i \int_0^t h_i(t, s)\, ds + r(t, \varepsilon) \right) \quad \text{with } |r(t, \varepsilon)| \leqq 2\varepsilon.$$

Now let us put

$$u_\varepsilon(t) = \begin{cases} v_i(t) & \text{if } t \in E_i, \\ u(t) & \text{if } t \notin \bigcup_{i=1}^{k} E_i. \end{cases}$$

Evidently, $u_\varepsilon \in \mathcal{U}_{\text{ad}}(0, T)$ by condition $(\mathcal{U})$; moreover, by (4.2) and (4.3) we have

$$\begin{aligned} &\int_0^t \mathbf{G}(t-s)[f(s, x(s), u_\varepsilon(s)) - f(s, x(s), u(s))]\, ds \\ &= \varepsilon \left(\sum_{i=1}^{k} \lambda_i \int_0^t \mathbf{G}(t-s)[f(s, x(s), v_i(s)) - f(s, x(s), u(s))\, ds] + r(t, \varepsilon) \right) \\ &= \varepsilon \left(\sum_{i=1}^{k} \lambda_i \left[y^i - \int_0^t \mathbf{G}(t-s)\mathbf{B}(s) y^i(s)\, ds \right] + r(t, \varepsilon) \right) \\ &= \varepsilon \left(y(t) - \int_0^t \mathbf{G}(t-s)\mathbf{B}(s) y(s)\, ds + r(t, \varepsilon) \right). \end{aligned} \tag{4.4}$$

Now the proof of the assertion can be split into the following two steps:

Step 1. For $\varepsilon > 0$ sufficiently small the Volterra integral equation

$$(4.5)\qquad x(t) = \mathbf{G}(t)x^0 + \int_0^t \mathbf{G}(t-s) f(s, x(s), u_\varepsilon(s))\, ds, \qquad t \in [0, T]$$

(and therefore problem (2.2)) has a solution $x_\varepsilon = (x_{\varepsilon 1}, x_{\varepsilon 2}) \in C(0, T; X)$; moreover, $x_\varepsilon \to x$ as $\varepsilon \to 0$ and $z_\varepsilon(t) = (x_\varepsilon(t) - x(t))/\varepsilon$ is uniformly bounded with respect to ε and t.

Step 2. z_ε is uniformly convergent to y as $\varepsilon \to 0$.

Proof of Step 1. By using a compactness argument it is easy to deduce from (f.3) that there exist $r > 0$, $L : [0, T] \times U \to \mathbf{R}_+$ and $\varphi : [0, r] \to \mathbf{R}_+$ satisfying (f.3)(1), (f.3)(2) and

$$\|f(t, y(t), u) - f(t, z(t), u)\| \leqq L(t, u)|y(t) - z(t)|,$$

$$\|(f_x(t, y(t), u) - f_x(t, x(t), u))(y(t) - x(t))\| \leqq L(t, u)\varphi(|y(t) - x(t)|)$$

for all $t \in [0, T]$, $u \in U$, y, z in the ball with center x and radius r in $C(0, T; X)$. From this and from (f.2)(3) it follows that there exist $r > 0$, $\varphi : [0, r] \to \mathbf{R}_+$ satisfying (f.3)(2), $\hat{L} = L(\cdot, u(\cdot)) + \sum_{i=1}^{k} L(\cdot, v_i(\cdot)) \in L^2(0, T)$ and $M \in L^q(0, T)$, $q > 2$ such that

$$(4.6)\quad \|\mathbf{B}(t)\| = \|f_x(t, x(t), u(t))\| \leqq M(t), \quad \|f_x(t, x(t), v_i(t))\| \leqq M(t) \quad (i = 1, \cdots, k),$$

$$(4.7)\qquad \|f(t, y(t), u_\varepsilon(t)) - f(t, z(t), u_\varepsilon(t))\| \leqq \hat{L}(t)|y(t) - z(t)|,$$

$$(4.8)\quad \|(f_x(t, y(t), u_\varepsilon(t)) - f_x(t, x(t), u_\varepsilon(t)))(y(t) - x(t))\| \leqq \hat{L}(t)\varphi(|y(t) - x(t)|)$$

for all $t \in [0, T]$ and for all y, z in the ball with center x and radius r in $C(0, T; X)$.

On the other hand, by (3.5) we have for some $c_1 > 0$

$$\left| \int_0^t \mathbf{G}(t-s)z(s)\, ds \right| \leqq c_1 \left(\int_0^t \|z(s)\|^2\, ds \right)^{1/2}, \tag{4.9}$$

for all $t \in [0, T]$, $z = (z_1, z_2) \in L^2(0, T; H' \times X_2)$.

Now, for all $\varepsilon > 0$, t, n, let us put

$$x_\varepsilon^0(t) = x(t) = \mathbf{G}(t)x^0 + \int_0^t \mathbf{G}(t-s)f(s, x(s), u(s))\, ds,$$

$$x_\varepsilon^{n+1}(t) = \mathbf{G}(t)x^0 + \int_0^t \mathbf{G}(t-s)f(s, x_\varepsilon^n(s), u_\varepsilon(s))\, ds.$$

By (4.4), (f.3)(1), (4.6), and (4.9), for all $t \in [0, T]$, $\varepsilon < 1$, we have

$$\begin{aligned} |x_\varepsilon^1(t) - x_\varepsilon^0(t)| &= \left| \int_0^t \mathbf{G}(t-s)[f(s, x(s), u_\varepsilon(s)) - f(s, x(s), u(s))]\, ds \right| \\ &\leqq \varepsilon \left[|y(t)| + \left| \int_0^t \mathbf{G}(t-s)\mathbf{B}(s)y(s)\, ds \right| + |r(t, \varepsilon)| \right] \\ &\leqq \varepsilon \left(\sup_{0 \leqq t \leqq T} |y(t)| + c_1 \left[\int_0^T \|\mathbf{B}(s)y(s)\|^2\, ds \right]^{1/2} + 2 \right) = \varepsilon c_2. \end{aligned}$$

Therefore, if we put $c = c_1 \|\hat{L}\|_2$ and $\varepsilon_0 = \min(1, r/c_2 e^c)$, then for all $\varepsilon < \varepsilon_0$ we have

$$\|x_\varepsilon^1 - x\| = \sup_{0 \leqq t \leqq T} |x_\varepsilon^1(t) - x_\varepsilon^0(t)| \leqq \varepsilon c_2 \leqq r e^{-c}.$$

Now by induction it is easy to see that for such $\varepsilon > 0$ we have (by (4.7) and (4.9))

$$|x_\varepsilon^n(t) - x(t)| \leqq \sum_{j=0}^{n-1} \frac{c^j}{j!} \|x_\varepsilon^1 - x\| \leqq r,$$

$$|x_\varepsilon^n(t) - x_\varepsilon^{n+1}(t)| \leqq \frac{c^n}{n!} \|x_\varepsilon^1 - x\|.$$

This proves that the sequence $(x_\varepsilon^n)_n$ converges to some x_ε in $C(0, T; X)$, with $\|x_\varepsilon(t) - x(t)\| \leqq e^c \|x_\varepsilon^1 - x\| \leqq \varepsilon c_2 e^c \leqq r$, for all $t \in [0, T]$, $\varepsilon < \varepsilon_0$. From this it follows that $x_\varepsilon \to x$ in $C(0, T; X)$ as $\varepsilon \to 0$, that

$$|z_\varepsilon(t)| = |(x_\varepsilon(t) - x(t))/\varepsilon| \leqq c_3 = c_2 e^c, \tag{4.10}$$

and that x_ε is a solution of the Volterra integral equation (4.5), since by (4.7) and (4.9) we have

$$\begin{aligned} &\left| \int_0^t \mathbf{G}(t-s)[f(s, x_\varepsilon^n(s), u_\varepsilon(s)) - f(s, x_\varepsilon(s), u_\varepsilon(s))]\, ds \right| \\ &\leqq c_1 \left(\int_0^t (\hat{L}(s)|x_\varepsilon^n(s) - x_\varepsilon(s)|)^2\, ds \right)^{1/2}. \end{aligned}$$

Proof of Step 2. By (4.4) and the fact that

$$f(s, x_\varepsilon(s), u_\varepsilon(s)) - f(s, x(s), u_\varepsilon(s)) = \int_0^1 f_x(s, x(s) + \tau\varepsilon z_\varepsilon(s), u_\varepsilon(s))(\varepsilon z_\varepsilon(s))\, d\tau,$$

we have that

$$z_\varepsilon(t)-y(t)=\frac{1}{\varepsilon}\int_0^t \mathbf{G}(t-s)[f(s, x_\varepsilon(s), u_\varepsilon(s))-f(s, x(s), u_\varepsilon(s))]\,ds$$

$$+\frac{1}{\varepsilon}\int_0^t \mathbf{G}(t-s)[f(s, x(s), u_\varepsilon(s))-f(s, x(s), u(s))]\,ds-y(t)$$

$$(4.11)\qquad =\int_0^t\int_0^1 \mathbf{G}(t-s)[f_x(s, x(s)+\tau\varepsilon z_\varepsilon(s), u_\varepsilon(s))-f_x(s, x(s), u_\varepsilon(s))]z_\varepsilon(s)\,d\tau\,ds$$

$$+\int_0^t \mathbf{G}(t-s)[f_x(s, x(s), u_\varepsilon(s))-f_x(s, x(s), u(s))]z_\varepsilon(s)\,ds$$

$$+\int_0^t \mathbf{G}(t-s)\mathbf{B}(s)(z_\varepsilon(s)-y(s))\,ds+r(t, \varepsilon).$$

Now the first addend of the right-hand side of (4.11) tends to zero uniformly with respect to t. In fact, for $\sigma>0$ fixed, by (f.3)(2) and (4.10) there exists $\varepsilon_1>0$ such that for all $\varepsilon<\varepsilon_1$, $s\in[0, T]$, $\tau\in[0, 1]$ we have

$$\|[f_x(s, x(s)+\tau\varepsilon z_\varepsilon(s), u_\varepsilon(s))-f_x(s, x(s), u_\varepsilon(s))]z_\varepsilon(s)\|\leqq \hat{L}(s)\varphi(|\tau\varepsilon z_\varepsilon(s)|)/\tau\varepsilon$$
$$\leqq \hat{L}(s)\sigma|z_\varepsilon(s)|\leqq \sigma c_3\hat{L}(s).$$

From this by (f.2)(2) and (4.9) it follows that for all $\varepsilon<\varepsilon_1$, $t\in[0, T]$ the norm of the first addend of the last term of (4.11) is bounded by $\sigma c_1c_3\|\hat{L}\|_2$.

On the other hand, by (4.6) and (4.9) the second addend of the right-hand side of (4.11) is bounded by

$$c_1\left(\int_E (2M(s)|z_\varepsilon(s)|)^2\,ds\right)^{1/2}\leqq 2c_1c_3\left(\int_E M(s)^2\,ds\right)^{1/2},$$

which tends to zero uniformly with respect to t as $\varepsilon\to 0$ since meas $E=\varepsilon T$.

Finally, the third addend of the right-hand side of (4.11) is bounded by

$$c_1\left(\int_0^t (M(s)|z_\varepsilon(s)-y(s)|)^2\,ds\right)^{1/2}.$$

Hence by (4.11) there exists $\sigma=\sigma(\varepsilon)>0$ such that $\sigma(\varepsilon)\to 0$ as $\varepsilon\to 0$ and

$$|z_\varepsilon(t)-y(t)|^2\leqq c_1^2\int_0^t (M(s)|z_\varepsilon(s)-y(s)|)^2\,ds+\sigma(\varepsilon).$$

From this and Remark 3.5 it follows that

$$|z_\varepsilon(t)-y(t)|^2\leqq \sigma(\varepsilon)\exp\left(c_1^2\int_0^t (M(s))^2\,ds\right),$$

which proves the assertion.

Remark 4.2. If f_1 maps $[0, T]\times X_1\times X_2\times U$ into X_1 rather than into H', then we use the estimate $|\mathbf{G}(t)x|\leqq|x_1|+|x_2|$ rather than (4.9), and easily prove the assertion of Lemma 4.1 under the weaker assumption that (f.2) holds with $q=1$.

Remark 4.3. Assume that there exist two Banach spaces X_1^0, X_2^0 that are densely embedded in X_1, X_2, respectively, and $f=(f_1, f_2)$ is a mapping from $[0, +\infty]\times X_1^0\times X_2^0\times U$ into $H'\times X_2$ satisfying conditions (f.1)-(f.3) with X_1, X_2 replaced by X_1^0, X_2^0. Moreover, assume that $f_x(t, x, u)$ is continuous from $X^0=X_1^0\times X_2^0$ into $H'\times X_2$ for the norm of X for all (t, x, u), and that $x(t)\in X^0$ almost everywhere in $[0, T]$ for any solution (x, u) of (2.2).

Then the proof of Lemma 4.1 can be repeated word for word, by denoting by $f_x(t, x(t), u(t))$ the linear continuous extension to X of the Frechét derivative of f.

5. The proof of the theorems. First, note that from Remark 3.3, (with A and $B(t)$ replaced by $\mathbf{A}$ and $\mathbf{B}(t)$), it follows that for all $\xi=(\xi_1, \xi_2)\in X_1\times X_2$ the adjoint equations (2.5) with final conditions $p(T)=\xi$ have a unique solution (p_1, p_2) with $p_2\in AC(0, T; X_2)$, $p_1\in C(0, T; X_1)\cap L^2(0, T; H)$, and $p_1-G(\cdot)\xi_1\in W^{1,q^*}(0, T; X_1)$, $q^*=2q/(2+q)$. Actually, $p_1\in W^{1,2}(0, T; X_1)\cap C(0, T; H)$ if $\xi_1\in H$ and $p_1\in W(0, T; H)$ if (f.2)(3) holds for $q=+\infty$.

Now we prove Theorems 2.1-2.3.

Proof of Theorems 2.1-2.3. Let us denote by $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ the subsets of $\mathbf{R}^2\times X$ defined by

$$\mathbf{\Sigma}=\{(\alpha, \lambda, y)\,|\,\lambda\leqq 0,\ y+x(T)\in B\},\qquad \mathbf{\Lambda}=\text{convex hull}\,(\Lambda_1+\Lambda_2),$$

where $\Lambda_1=\{(\alpha, \lambda, y)\,|\,\alpha=0,\ \lambda=\Phi_x(T, x(T))y,\ y=y^v(T)$ for some $v\in\mathcal{U}_{\text{ad}}(0, T)\}$ and $\Lambda_2=\Gamma\cdot(\bar\alpha, \bar\lambda, \bar y_1, \bar y_2)$ with

$$\begin{aligned}
&\Gamma=\mathbf{R},\quad \bar\alpha=0,\quad \bar\lambda=0,\quad \bar y_1=0,\quad \bar y_2=0 \quad \text{for Theorem 2.1,}\\
&\Gamma=\mathbf{R}_-,\quad \bar\alpha=1,\quad \bar\lambda=\mathbf{\Phi}_t(T, x(T))+\mathbf{\Phi}_{x_2}(T, x(T))\bar y_2,\\
&\qquad\qquad \bar y_1=0,\quad \bar y_2=f_2^-(T, x(T), u(T)) \quad \text{for Theorem 2.2,}\\
&\Gamma=\mathbf{R},\quad \bar\alpha=1,\quad \bar\lambda=1,\quad \bar y_1=0,\quad \bar y_2=f_2^-(T, x(T), u(T)) \quad \text{for Theorem 2.3.}
\end{aligned}\tag{5.1}$$

Evidently, $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ are convex subsets of $\mathbf{R}^2\times X$ and $0\in\mathbf{\Sigma}\cap\mathbf{\Lambda}$, since $x(T)\in B$ and $y^u(T)=0$. Moreover, $\mathbf{\Sigma}$ has nonempty interior (actually, $(\alpha, \lambda, y)\in\text{int}\,(\mathbf{\Sigma})$ if and only if $\lambda<0$, and $y+x(T)\in\text{int}\,(B)$). Let us prove that

$$\mathbf{\Lambda}\cap\text{int}\,(\mathbf{\Sigma})=\varnothing.\tag{5.2}$$

In fact, assume by contradiction that there exists $(\alpha, \lambda, y)\in\mathbf{\Lambda}\cap\text{int}\,(\mathbf{\Sigma})$. Then there exist $\gamma_1, \cdots, \gamma_k\in\Gamma$, $v_1, \cdots, v_k\in\mathcal{U}_{\text{ad}}(0, T)$, and $\lambda_1, \cdots, \lambda_k\in\mathbf{R}_+$ with $\sum_{i=1}^k\lambda_i=1$, such that

$$\alpha=\bar\alpha\sum_{i=1}^k\gamma_i\lambda_i,\tag{5.3}$$

$$\lambda=\sum_{i=1}^k\lambda_i\Phi_x(T, x(T))y^{v_i}(T)+\alpha\bar\lambda\quad\text{and}\quad\lambda<0,\tag{5.4}$$

$$y=\sum_{i=1}^k\lambda_i y^{v_i}(T)+\alpha\bar y\quad\text{with } y+x(T)\in\text{int}\,(B).\tag{5.5}$$

(Note that by (5.1), (5.3), and (5.4), we have $\lambda=\Phi_x(T, x(T))y+\alpha\Phi_t(T, x(T))$ and $\alpha\leqq 0$ in all cases.)

Now by Lemma 4.1 there exists $\varepsilon_0>0$ such that for all $\varepsilon\in]0, \varepsilon_0]$ there exists a solution $(x_\varepsilon, u_\varepsilon)$ of (2.2) such that $(x_\varepsilon-x)/\varepsilon\to\sum_{i=1}^k\lambda_i y^{v_i}$ uniformly in $[0, T]$ as $\varepsilon\to 0^+$. From this and (5.5) it follows that

$$\begin{aligned}
&\lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}(x_\varepsilon(T)-x(T))=y \quad \text{in the case of Theorem 2.1,}\\
&\lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}(x_{\varepsilon 2}(T+\varepsilon\alpha)-x_2(T))=y_2 \quad \text{in the case of Theorems 2.2 and 2.3,}
\end{aligned}\tag{5.6}$$

since we have

$$\lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}(x_{\varepsilon 2}(T+\varepsilon\alpha)-x_2(T+\varepsilon\alpha))-\sum_{i=1}^{k}\lambda_i y_2^{v_i}(T+\varepsilon\alpha)=0,$$

$$\lim_{\varepsilon\to 0^+}\sum_{i=1}^{k}\lambda_i y_2^{v_i}(T+\varepsilon\alpha)=\sum_{i=1}^{k}\lambda_i y_2^{v_i}(T),$$

$$\lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}(x_2(T+\varepsilon\alpha)-x_2(T))=\lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\int_T^{T+\varepsilon\alpha} f_2(s,x(s),u(s))\,ds=\alpha f_2^-(T,x(T),u(T)).$$

From (5.6) and (5.5), by the convexity of B (respectively, B_2), it follows that (for all $\varepsilon>0$ sufficiently small), $x_\varepsilon(T)=(1-\varepsilon)x(T)+\varepsilon(x(T)+y+o(1))\in B$ (respectively, $x_{\varepsilon 2}(T+\alpha\varepsilon)=(1-\varepsilon)x_2(T)+\varepsilon(x_2(T)+y_2+o(1))\in B_2$).

Hence $(T, x_\varepsilon, u_\varepsilon)$ (respectively, $(T+\varepsilon\alpha, x_\varepsilon, u_\varepsilon)$) satisfies the final state constraint (2.3); therefore by the optimality of (T, x, u) we have $\Phi(T, x(T))\leqq\Phi(T, x_\varepsilon(T))$ (respectively, $\Phi(T, x(T))\leqq\Phi(T+\varepsilon\alpha, x_\varepsilon(T+\varepsilon\alpha))$). From this, by (5.6) and the differentiability of Φ, it follows that $0\leqq\lim_{\varepsilon\to 0^+}(\Phi(T, x_\varepsilon(T))-\Phi(T, x(T)))/\varepsilon=\Phi_x(T, x(T))(y)=\lambda$, (respectively, $0\leqq\lim_{\varepsilon\to 0^+}1/\varepsilon(\Phi(T+\varepsilon\alpha, x_\varepsilon(T+\varepsilon\alpha)-\Phi(T, x_\varepsilon(T)))=\alpha\Phi_t(T, x(T))+\Phi_x(T, x(T))(y_2)=\lambda$), contradicting (5.4).

Hence Λ and Σ are convex and (5.2) holds. From this, by the Eidelheit separation theorem, it follows that there exist $(\alpha_0, \lambda_0, x^*)\in\mathbf{R}^2\times X$, $x^*=(x_1^*, x_2^*)$ such that

$$(\alpha_0, \lambda_0, x^*)\neq 0, \tag{5.7}$$

$$\alpha_0\alpha+\lambda_0\lambda+(x^*, y)\geqq 0 \quad \text{in } \Lambda, \tag{5.8}$$

$$\alpha_0\alpha+\lambda_0\lambda+(x^*, y)\leqq 0 \quad \text{in } \Sigma. \tag{5.9}$$

From (5.9) (for $\alpha\in\mathbf{R}$, $\lambda=0$, $y=0$ and for $\alpha=0$, $\lambda=-1$, $y=0$) it follows that $\alpha_0=0$ and $\lambda_0\geqq 0$; therefore we can assume that (5.7)-(5.9) hold for $\alpha_0=0$, $\lambda_0\in\{0, 1\}$.

Now let $p=(p_1, p_2)$ be the unique solution of the adjoint equations (2.5) with final conditions $p(T)=x^*+\lambda_0\Phi_x(T, x(T))$. Then by (5.7) we have the nondegeneracy condition (2.4) and by (5.9) (for $\lambda=0$) we have the transversality conditions (2.7) (for Theorem 2.1) or (2.7a) (for Theorems 2.2 and 2.3).

On the other hand, from (5.8) it follows that $\lambda_0(\lambda-\Phi_x(T, x(T))y)+(p(T), y)\geqq 0$ for all $(\alpha, \lambda, y)\in\Lambda_1\cup\Lambda_2$, i.e.,

$$(p(T), y^v(T))\geqq 0 \quad \text{for all } v\in\mathcal{U}_{\text{ad}}(0, T), \tag{5.10}$$

$$\gamma[\lambda_0\bar{\lambda}-\Phi_x(T, x(T))\bar{y})+(p(t), \bar{y})]\geqq 0 \quad \text{for all } \gamma\in\Gamma. \tag{5.11}$$

Then the transversality condition (2.7b) or (2.7b′) follows from (5.11) and the definition of Γ, $\bar{\lambda}$, $\bar{y}$. Finally, the maximum principle (2.6) follows from (5.10), since by Remark 3.4 (with A and B replaced by $\mathbf{A}$ and $\mathbf{B}$), we have

$$\begin{aligned}(p(T), y^v(T)) &= \int_0^T \langle f(s, x(s), v(s))-f(s, x(s), u(s)), p(s)\rangle\,ds\\ &= \int_0^T (H(s, x(s), v(s), p(s))-H(s, x(s), u(s), p(s))\,ds.\end{aligned}$$

Remark 5.1. Evidently, a central role in the proof of Theorems 2.1-2.3 is played by Lemma 4.1. Hence the assertions of Theorems 2.1-2.3 hold under the assumptions of Remark 4.2 or Remark 4.3.

Remark 5.2. Under the assumptions of Theorem 2.2, if we replace condition (3) with the stronger assumption that (x, u) can be prolonged to an interval $[0, T']$, $T' > T$ in such a way that there exists $\lim_{t\to T} f_2(t, x(t), u(t))$, then we have the equality sign in (2.7b).

In fact we have only to repeat word for word the above proof by replacing $[0, T]$ with $[0, T']$, and $\Gamma = \mathbf{R}_-$ with $\Gamma = \mathbf{R}$ in the definition of Λ_2.

Proof of Corollaries 2.4 *and* 2.5. Evidently, $(T, \hat{x}, u)$ with $\hat{x} = (x_1, \cdots, x_{l+2})$ and $x_{j+2}(t) = \int_0^t h_j(u(s))\, ds$ for all $j = 1, \cdots, l$, is an optimal solution of the following modified problem.

Minimize $J(T, \hat{x}, u) = \hat{\Phi}(T, \hat{x}(T)) = \Phi(T, x_1(T), x_2(T))$ with $\hat{x} = (x_1, \cdots, x_{l+2})$, u satisfying with (2.2) the additional state equations $x'_{j+2} = h_j(u(\cdot))$, $x_{j+2}(0) = 0$, for all j, and the final constraint $\hat{x}(T) \in \hat{B} = X_1 \times B_2 \times]-\infty, c_1] \times \cdots \times]-\infty, c_l]$.

Now for all $v \in \mathcal{U}_{\text{ad}}(0, T)$ let $\hat{y}^v = (y^1, \cdots, y_{l+2})$ be the unique solution of (4.1) and $y'_{j+2} = h_j(v(\cdot)) - h_j(u(\cdot))$, $y_{j+2}(0) = 0$, for all $j = 1, \cdots, l$, and let us put

$$\hat{\mathbf{\Sigma}} = \{(\alpha, \lambda, \hat{y}) \,|\, \lambda \leqq 0, \hat{y} + \hat{x}(T) \in \hat{B}\}, \qquad \hat{\mathbf{\Lambda}} = \text{convex hull } (\hat{\Lambda}_1 + \hat{\Lambda}_2)$$

where $\hat{\Lambda}_1 = \{(\alpha, \lambda, \hat{y}) \,|\, \alpha = 0,\ \lambda = \hat{\Phi}_x(T, \hat{x}(T))\hat{y},\ \hat{y} = \hat{y}^v(T) \text{ for some } v \in \mathcal{U}_{\text{ad}}(0, T)\}$ and $\hat{\Lambda}_2 = \Gamma \cdot (\bar{\alpha}, \bar{\lambda}, \bar{y}_1, \bar{y}_2, \cdots, \bar{y}_{l+2})$ with $\Gamma, \bar{\alpha}, \bar{\lambda}, \bar{y}_1, \bar{y}_2$ defined in (5.1) and $\bar{y}_j = 0$ for $j > 2$.

Then the proof parallels the previous proof. In particular, (5.5) becomes

$$\sum_{i=1}^{k} \lambda_i y_2^{v_i}(T) + \alpha f_2^-(T, x(T), u(T)) + x_2(T) \in \text{int } (B_2),$$

$$\sum_{i=1}^{k} \lambda_i y_{j+2}^{v_i}(T) + x_{j+2}(T) < c_j \quad \text{for all } j = 1, \cdots, l.$$

Moreover, we have that $(\hat{x}_\varepsilon - \hat{x})/\varepsilon \to \sum_{i=1}^{k} \lambda_i \hat{y}^{v_i}$ uniformly in $[0, T]$. From this it follows that $x_{\varepsilon 2}(T + \varepsilon\alpha) \in B_2$ with the same argument used before, whereas from the fact that $\alpha \leqq 0$ and $h_j(z) \geqq 0$ for all z, j we deduce that

$$x_{\varepsilon, j+2}(T + \varepsilon\alpha) = \int_0^{T+\varepsilon\alpha} h_j(u_\varepsilon(s))\, ds \leqq \int_0^T h_j(u_\varepsilon(s))\, ds = x_{\varepsilon, j+2}(T)$$

$$= (1 - \varepsilon) x_{j+2}(T) + \varepsilon \left(c_{j+2}(T) + \sum_{i=1}^{k} \lambda_i y_{j+2}^{v_i}(T) + o(1) \right) < c_j.$$

Hence $\hat{x}_\varepsilon$ satisfies the constraint $\hat{x}_\varepsilon(T + \varepsilon\alpha) \in \hat{B}$ and the assertion follows as before.

Remark 5.3. If we replace the nonnegativity assumption $h_j(z) \geqq 0$ for all z, j with the assumption that there exists $\lim_{t\to T^-} h_j(u(T)) = h_j^-(u(t))$ for all j, then the assertion of Corollary 2.4 holds with (2.7b) replaced by

$$H(T, x(T), u(T), p(T)) + \sum_{j=1}^{l} \mu_j h_j^-(u(T)) \leqq -\lambda_0 \Phi_t(T, x(T)).$$

Moreover, in the last condition we have the equality sign if $\Phi(t, x) = t$ or if (x, u) can be prolonged to an interval $[0, T']$, $T' > T$, in such a way that there exist $\lim_{t\to T} f_2(t, x(t), u(t))$, and $\lim_{t\to T} h_j(u(t))$ for all $j = 1, \cdots, l$.

In fact, we need only to apply Theorems 2.2 and 2.3 to the modified optimal control problem described in the proof of Corollaries 2.4 and 2.5.

Remark 5.4. Note that the assertion of Theorem 2.3 or Corollary 2.5 holds in the case when $\Phi(t, x) = \gamma(t)$, with $\gamma : \mathbf{R}_+ \to \mathbf{R}$ increasing and left differentiable. We have only to replace $\bar{\lambda} = 1$ with $\bar{\lambda} = \gamma'_-(T)$, in the definition of Λ_2.

6. Two applications to the study of the diffusion of a class of epidemics.

6.1. First application. Let Ω be an open bounded subset of $\mathbf{R}^2$ whose boundary $\partial\Omega$ is sufficiently smooth and is the union of two disjoint curves Γ_1, Γ_2 and let us consider the problem

$$\frac{\partial y_1}{\partial t}(t, x) - \Delta y_1(t, x) + a_1 y_1(t, x) = \omega_1(t, x) \int_\Omega k_1(x, \xi) y_2(t, \xi)\, d\xi \quad \text{in } Q^T,$$

$$\frac{\partial y_2}{\partial t}(t, x) + a_2 y_2(t, x) = g(y_1(t, x)) \quad \text{in } Q^T,$$

$$\frac{\partial y_1}{\partial \nu}(t, \sigma) + \alpha y_1(t, \sigma) = \omega_2(t, \sigma) \int_\Omega k_2(\sigma, \xi) y_2(t, \xi)\, d\xi \quad \text{on } \Sigma_1^T, \tag{6.1}$$

$$\frac{\partial y_1}{\partial \nu}(t, \sigma) = 0 \quad \text{on } \Sigma_2^T,$$

$$y_1(0, x) = y_1^0(x), \quad y_2(0, x) = y_2^0(x) \quad \text{in } \Omega$$

where $Q^T = [0, T] \times \Omega$, $\Sigma_1^T = [0, T] \times \Gamma_1$, and $\Sigma_2^T = [0, T] \times \Gamma_2$.

This problem describes the diffusion of an epidemic of oro-fecal origin (such as cholera, typhoid fever, and so on) that has been exhaustively studied by Capasso and his co-workers (see [1], [7], [8] and the references therein). More precisely, $y_1(t, x)$ and $y_2(t, x)$ represent, respectively, the density of an infectious agent and infected persons at the time t in the point x; the function $g = g(y_1)$ represents the strength of the contagion, the Laplace operator Δ represents the random diffusion of the infective agent in the habitat, whereas the integral operators

$$\int_\Omega k_1(x, \xi) y_2(t, \xi)\, d\xi \quad \text{and} \quad \int_\Omega k_2(\sigma, \xi) y_2(t, \xi)\, d\xi$$

represent the diffusive effects of the epidemic produced by the infected persons; finally, the functions $\omega_1 = \omega_1(t, x)$ and $\omega_2 = \omega_2(t, \sigma)$ are two factors of reductions of the diffusive effects of the epidemics that are produced by suitable sanitation programs in the habitat Ω and on Γ_1. For a complete description of such a model see [1] and [5], where it is also proved that for every $T > 0$, $\omega_1 \in L^\infty(Q^T) \simeq L^\infty(0, T; L^\infty(\Omega))$, $\omega_2 \in L^\infty(\Sigma_1^T) \simeq L^\infty(0, T; L^\infty(\Gamma_1))$, and for every $y_1^0, y_2^0 \in L^\infty(\Omega)$, there exists a unique weak solution (y_1, y_2) of (6.1) with

$$y_1 \in W(0, T; H^1(\Omega)) \cap C(0, T; L^2(\Omega)) \cap L^\infty(Q^T),$$

$$y_2 \in AC(0, T, L^\infty(\Omega)) \cap C^1(0, T; L^2(\Omega)).$$

Now the purpose of the public authorities is to choose a sanitary strategy through a fixed time interval $[0, T]$ that allows them to "win the diffusion of epidemic," in the sense that at the time T the total infected population must be sufficiently small, i.e.,

$$\int_\Omega y_2(T, x)\, dx \leqq \bar{y}_2 \quad \text{with } \bar{y}_2 \text{ a preassigned positive constant.} \tag{6.2}$$

Finally, connected with the epidemic there is a cost depending on the infected population that has the form

$$\int_0^T \int_\Omega f(y_2(t,x))\,dx\,dt,$$

and there is a cost of the sanitation program that has the form

$$\int_0^T \left[\int_\Omega h_1(\omega_1(t,x))\,dx + \int_{\Gamma_1} h_2(\omega_2(t,\sigma))\,d\sigma \right] dt.$$

Obviously the purpose of the public authorities is to choose an optimal strategy, i.e., a sanitary strategy (producing the optimal controls ω_1, ω_2) in such a way that the corresponding evolution of the epidemic (y_1, y_2) satisfies the final condition (5.2) and minimizes the cost functional

$$J(y_1, y_2, \omega_1, \omega_2) = \int_0^T \left[\int_\Omega f(y_2(t,x))\,dx + \int_\Omega h_1(\omega_1(t,x))\,dx + \int_{\Gamma_1} h_2(\omega_2(t,\sigma))\,d\sigma \right] dt.$$

Then we have the following theorem.

THEOREM 6.1. *Assume that*

$$k_1 \in L^\infty(\Omega\times\Omega), \quad k_2 \in L^\infty(\Gamma_1\times\Omega), \quad k_1 \geqq 0, \quad k_2 \geqq 0,$$

$$f, g \in C^1(\mathbf{R}, \mathbf{R}) \quad \textit{with } |g(z)| \leqq c(1+|z|) \quad \textit{for some } c>0,$$

$$y_1^0, y_2^0 \in L^\infty(\Omega), \quad y_1^0 \geqq 0, \quad y_2^0 \geqq 0,$$

$$h_1, h_2 : [0,1] \to [0,+\infty] \quad \textit{convex decreasing functions.}$$

Moreover, assume that $\omega_1 \in L^\infty(Q^T)$, $\omega_2 \in L^\infty(\Sigma_1^T)$ *are optimal controls in the sense that* ω_1, ω_2 *and the corresponding solution* y_1, y_2 *of* (6.1) *do minimize the functional* J *among all the weak solutions* $(y_1, y_2, \omega_1, \omega_2)$ *of* (6.1) *such that* $\int_\Omega y_2(T,x)\,dx \leqq \bar{y}_2$. *Then there exist* $\lambda_0 \in \{0,1\}$, $\mu_0 \geqq 0$, $p_1 \in C(0,T; H^1(\Omega)) \cap W^{1,2}(0,T; L^2(\Omega))$, $p_2 \in AC(0,T; L^2(\Omega))$ *such that* $(\lambda_0, \mu_0) \neq (0,0)$ *and* p_1, p_2 *satisfy in* Q^T *the adjoint equations*

$$\begin{aligned} &\frac{\partial p_1}{\partial t}(t,x) - a_1 p_1(t,x) + \Delta p_1(t,x) = -g'(y_1(t,x))p_2(t,x), \\ &\frac{\partial p_2}{\partial t}(t,x) - a_2 p_2(t,x) = -\int_\Omega k_1(\xi,x)\omega_1(t,\xi)p_1(t,\xi)\,d\xi \\ &\qquad\qquad - \int_{\Gamma_1} k_2(\sigma,x)\omega_2(t,\sigma)p_1(t,\sigma)\,d\sigma - \lambda_0 f'(y_2(t,x)), \end{aligned} \tag{6.3}$$

with boundary and final conditions

$$\frac{\partial p_1}{\partial \nu} + \alpha p_1 = 0 \quad \textit{on } \Sigma_1^T, \qquad \frac{\partial p_1}{\partial \nu} = 0 \quad \textit{on } \Sigma_2^T, \tag{6.4}$$

$$p_1(T) \equiv 0, \qquad p_2(T) \equiv \text{const.} = \mu_0 \quad \textit{and} \quad \mu_0 = 0 \quad \textit{if} \int_\Omega y_2(T,x)\,dx < \bar{y}_2, \tag{6.5}$$

and the maximum principle

$$
\begin{aligned}
&-p_1(t,x)\int_\Omega k_1(x,\xi)y_2(t,\xi)\,d\xi \in \lambda_0 \partial h_1(\omega_1(t,x)) \quad \text{a.e. in } Q^T,\\
&-p_1(t,\sigma)\int_\Omega k_2(\sigma,\xi)y_2(t,\xi)\,d\xi \in \lambda_0 \partial h_2(\omega_2(t,\sigma)) \quad \text{a.e. in } \Sigma_1^T
\end{aligned}
\tag{6.6}
$$

(*where* ∂h_i *denotes the subdifferential of* h_i).

Proof. Let us put $X^0 = L^\infty(\Omega)$, $X = L^2(\Omega)$, $H = H^1(\Omega)$ with the usual norms and inner products. Then evidently X^0 and H are densely embedded in X and $H \subset X \simeq X' \subset H'$. Moreover, let us denote by A the linear continuous self-adjoint coercive operator from H to H' defined by

$$
\langle A\varphi, \psi\rangle = \int_\Omega (\nabla\varphi\nabla\psi + a_1\varphi\psi)\,dx + \alpha\int_{\Gamma_1}\varphi\psi\,d\sigma.
$$

Finally, let $Z_1 = L^\infty(\Omega)$, $Z_2 = L^\infty(\Gamma_1)$, $Z = Z_1 \times Z_2$, and $U = U_1 \times U_2$ where we have put $U_i = \{\omega \in Z_i \mid 0 \leqq \omega(x) \leqq 1 \text{ a.e.}\}$ $(i = 1, 2)$ and let us denote by $\mathcal{U}_{ad}(0, T)$ the set of elements $(\omega_1, \omega_2) \in L^\infty(Q^T) \times L^\infty(\Sigma_1^T)$ such that $0 \leqq \omega_i(t,x) \leqq 1$ almost everywhere $(i = 1, 2)$ and the mappings $\int_\Omega h_1(\omega_1(\cdot, x))\,dx$, $\int_{\Gamma_1} h_2(\omega_2(\cdot,\sigma))\,d\sigma$ are summable.

Then (y_1, y_2, y_3) with y_3 defined by

$$
y_3(t) = \int_0^t \left(\int_\Omega f(y_2(t,x))\,dx + \int_\Omega h_1(\omega_1(t,x))\,dx + \int_{\Gamma_1} h_2(\omega_2(t,\sigma))\,d\sigma\right) ds
$$

is a solution of the following optimal control problem.

To minimize the functional $J(y_1, y_2, y_3, \omega_1, \omega_2) = \Phi(y_1(T), y_2(T), y_3(T))$ with $y_1 \in W(0, T; H) \cap C(0, T; X)$, $y_2 \in AC(0, T; X)$, $y_3 \in AC(0, T; \mathbf{R})$, $(\omega_1, \omega_2) \in \mathcal{U}_{ad}(0, T)$, and

$$
\begin{aligned}
&y_1'(t) + Ay_1(t) = F_1(t, y_1(t), y_2(t), y_3(t), \omega_1(t), \omega_2(t)) \quad \text{a.e. in } [0, T],\\
&y_2'(t) = F_2(t, y_1(t), y_2(t), y_3(t), \omega_1(t), \omega_2(t)) \quad \text{a.e. in } [0, T],\\
&y_3'(t) = F_3(t, y_1(t), y_2(t), y_3(t), \omega_1(t), \omega_2(t)) \quad \text{a.e. in } [0, T],\\
&y_1(0) = y_1^0, \quad y_2(0) = y_2^0, \quad y_3(0) = 0,\\
&(y_1(T), y_2(T), y_3(T)) \in B,
\end{aligned}
\tag{6.7}
$$

where B is the set of the elements (y_1, y_2, y_3) of $X \times X \times \mathbf{R}$ such that $\int_\Omega y_2(x)\,dx \leqq \bar{y}_2$ and Φ, F_1, F_2, F_3 are operators from $[0, +\infty[\times X^0 \times X^0 \times \mathbf{R}$ in $\mathbf{R}$ and from $[0, +\infty[\times X^0 \times X^0 \times \mathbf{R} \times U$ into $H', X, \mathbf{R}$, respectively, defined by

$$
\Phi(t, y_1, y_2, y_3) = y_3, \tag{6.8}
$$

$$
\begin{aligned}
\langle F_1(t, y_1, y_2, y_3, \omega_1, \omega_2), \varphi\rangle = &\int_{\Omega\times\Omega} \varphi(x)k_1(x,\xi)\omega_1(x)y_2(\xi)\,dx\,d\xi\\
&+ \int_{\Gamma_1\times\Omega} \varphi(\sigma)k_2(\sigma,\xi)\omega_2(\sigma)y_2(\xi)\,d\sigma\,d\xi,
\end{aligned}
\tag{6.9}
$$

$$
F_2(t, y_1, y_2, y_3, \omega_1, \omega_2) = -a_2y_2 + G(y_1), \tag{6.10}
$$

$$
\begin{aligned}
F_3(t, y_1, y_2, y_3, \omega_1, \omega_2) = &\int_\Omega f(y_2(t,x))\,dx + \int_\Omega h_1(\omega_1(t,x))\,dx\\
&+ \int_{\Gamma_1} h_2(\omega_2(t,\sigma))\,d\sigma
\end{aligned}
\tag{6.11}
$$

and G is the Nemytskij operator associated by g, i.e., $G(y)(x) = g(y(x))$.

Now it is evident that B is convex and has nonempty interior. Moreover, every solution $y=(y_1, y_2, y_3)$ of (6.7) is such that $y(t)\in X^0\times X^0\times \mathbf{R}$ almost everywhere in $[0, T]$.

Finally, it can be easily shown that $F=(F_1, F_2, F_3)$ satisfies conditions (f.1)-(f.3) (with $q=+\infty$ and X_1, X_2 replaced by X^0) and that for every (t, y_1, y_2, y_3, u) the Frechét derivative of F at (t, y_1, y_2, y_3, u) is continuous for the norm of $X\times X\times \mathbf{R}$. Hence if we put

$$H(t, y, \omega, p)=\langle F_1(t, y, \omega), p_1\rangle+(F_2(t, y, \omega), p_2)+p_3F_3(t, y, \omega)$$

for all $t\in[0, T]$, $y=(y_1, y_2, y_3)\in X^0\times X^0\times \mathbf{R}$, $p=(p_1, p_2, p_3)\in H\times X\times \mathbf{R}$, $\omega\in U$, then by Theorem 2.1 and Remark 4.3 there exist λ_0, p_1, p_2, p_3 satisfying almost everywhere in $[0, T]$ the adjoint equations

$$\begin{aligned} &p_1'(t)-Ap_1(t)=-H_{y_1}(t, y(t), \omega(t), p(t)),\\ &p_2'(t)=-H_{y_2}(t, y(t), \omega(t), p(t)),\\ &p_3'(t)=-H_{y_3}(t, y(t), \omega(t), p(t))=0, \end{aligned} \tag{6.12}$$

the transversality condition

$$(p(T)-\lambda_0\phi_y(T, y(T)), y-y(T))\leqq 0 \text{ for all } y\in B, \tag{6.13}$$

the maximum principle

$$\int_0^T H(t, y(t), \omega(t), p(t))\,dt\leqq\int_0^T H(t, y(t), \bar{\omega}(t), p(t))\,dt \tag{6.14}$$

for all $\bar{\omega}\in \mathcal{U}_{ad}(0, T)$, and the nondegeneracy condition

$$(\lambda_0, p(T))\neq 0. \tag{6.15}$$

From (6.13) and the last equality of (6.12), it follows that

$$p_1(T, x)\equiv 0,$$

$$p_2(T, x)\equiv \text{const.}=\mu_0\geqq 0 \quad\text{and}\quad \mu_0=0 \quad\text{if}\quad \int_\Omega y_2(T, x)\,dx<\bar{y}_2,$$

$$p_3(t)\equiv \text{const.}=p_3(T) \quad\text{and}\quad p_3(t)-\lambda_0=0,$$

which proves that $(\lambda_0, \mu_0)\neq 0$ and that (6.5) holds. Moreover, by (6.12) we easily have that p_1, p_2 are solutions of the adjoint equations (6.3) with boundary conditions (6.4) and $p_1\in C(0, T; H^1(\Omega))\cap W^{1,2}(0, T; L^2(\Omega))$, since $p_1(T)=0\in H$.

Finally, by (6.14) we have that ω_1 minimizes in $L^\infty(Q^T)$ the functional J_1+J_2 where

$$J_1(\omega)=\begin{cases}\lambda_0\displaystyle\int_{Q^T} h_1(\omega(t, x))\,dx\,dt & \text{if } 0\leqq\omega(t, x)\leqq 1\leqq \text{ a.e. and the integral is finite,}\\ +\infty & \text{otherwise,}\end{cases}$$

$$J_2(\omega)=\int_{Q^T}\left(\omega(t, x)p_1(t, x)\int_\Omega k_1(x, \xi)y_2(t, \xi)\,d\xi\right)dx\,dt.$$

Since evidently J_2 is linear continuous in $L^2(Q^T)$ and J_1 is convex and lower semicontinuous, we have that

$$-p_1(\cdot,\cdot)\int_\Omega k_1(\cdot,\xi)y_2(\cdot,\xi)\,d\xi = -J_2'(\omega_1)\in\partial J_1(\omega_1),$$

i.e., (see [3, Prop. 2.7, p. 102]),

$$-p_1(t,x)\int_\Omega k_1(x,\xi)y_2(t,\xi)\,d\xi\in\lambda_0\partial h_1(\omega_1(t,x)) \quad \text{a.e. in } Q^T.$$

With the same arguments we obtain

$$-p_1(t,\sigma)\int_\Omega k_2(\sigma,\xi)y_2(t,\xi)\,d\xi\in\lambda_0\partial h_2(\omega_2(t,\sigma)) \quad \text{a.e. in } \Sigma_1^T.$$

6.2. Second application. In the preceding discussion the final time T, i.e., the length of the epidemic, was pre-assigned and the aim was to choose the optimal strategy to win the diffusion of the epidemic within the time T with a minimum cost.

Now let us consider the case where the time T is unknown, and we wish to choose the strategy that allows us to win the diffusion of the epidemic in the minimum time among all the strategies (ω_1,ω_2) with "acceptable costs." In other words let us consider the following time optimal control problem.

Minimize the functional $J(T,y_1,y_2,\omega_1,\omega_2)=T$ with $(y_1,y_2,\omega_1,\omega_2)$ a solution of (6.1) such that

$$\int_\Omega y_2(T,x)\,dx\leqq\bar{y}_2, \tag{6.16}$$

$$\int_{Q^T} h_1(\omega_1(t,x))\,dx\,dt\leqq c_1,$$

$$\int_{\Sigma_1^T} h_2(\omega_2(t,x))\,dx\,dt\leqq c_2, \tag{6.17}$$

$$\int_{Q^T} f(y_2(t,x))\,dx\,dt\leqq c_3.$$

(Condition (6.16) means that the epidemic has been won at the time T, whereas conditions (6.17) mean that the costs are acceptable.)

Such a problem can be put into the following form.

Minimize the functional $J(T,y_1,y_2,y_3,\omega_1,\omega_2)=T$ where $(y_1,y_2,y_3,\omega_1,\omega_2)$ is a solution of (6.7), with

$$F_3(t,y_1,y_2,y_3,\omega_1,\omega_2)=\int_\Omega f(y_2(t,x))\,dx$$

and

$$B=X\times B',\ B'=\left\{(y_2,y_3)\times X\times\mathbf{R}\;\middle|\;\int_\Omega y_2(x)\,dx\leqq\bar{y}_2;\ y_3\leqq c_3\right\},$$

satisfying the isoperimetric constraints

$$\int_0^T\int_\Omega h_1(\omega_1(t,x))\,dx\,dt\leqq c_1,\qquad \int_0^T\int_{\Gamma_1} h_2(\omega_2(t,x))\,dx\,dt\leqq c_2.$$

Then we have the following theorem.

THEOREM 6.2. *If T is the optimal time, (ω_1, ω_2) is the optimal control, and (y_1, y_2) is the corresponding optimal trajectory, then there exist $\lambda_1, \mu_0, \mu_1, \mu_2, \mu_3$ such that*

$$\lambda_0 \in \{0, 1\} \quad \text{and} \quad (\lambda_0, \mu_0, \mu_1, \mu_2, \mu_3) \neq 0, \tag{6.18}$$

$$\begin{aligned}
&\mu_0 \geqq 0 \quad \text{and} \quad \mu_0 = 0 \quad \text{if} \int_\Omega y_2(T, x)\, dx < \bar{y}_2,\\
&\mu_1 \geqq 0 \quad \text{and} \quad \mu_1 = 0 \quad \text{if} \int_{Q^T} h_1(\omega_1(t, x))\, dx\, dt < c_1,\\
&\mu_2 \geqq 0 \quad \text{and} \quad \mu_2 = 0 \quad \text{if} \int_{\Sigma_1^T} h_2(\omega_2(t, x))\, dx\, dt < c_2,\\
&\mu_3 \geqq 0 \quad \text{and} \quad \mu_3 = 0 \quad \text{if} \int_{Q^T} f(y_2(t, x))\, dx\, dt < c_3,
\end{aligned} \tag{6.19}$$

$$\lambda_0 + \mu_0 \int_\Omega [g(y_1(T, x)) - a_2 y_2(T, x)]\, dx + \mu_3 \int_\Omega f(y_2(T, x))\, dx = 0, \tag{6.20}$$

and there exist $p_1 \in C(0, T; H^1(\Omega)) \cap W^{1,2}(0, T; L^2(\Omega))$, $p_2 \in AC(0, T; L^2(\Omega))$, which are solutions of the adjoint equations (6.3) *with boundary and final conditions* (6.4), (6.5) *and satisfy conditions*

$$\begin{aligned}
&-p_1(t, x) \textstyle\int_\Omega k_1(x, \xi) y_2(t, \xi)\, d\xi \in \mu_1 \partial h_1(\omega_1(t, x)) \quad \text{a.e. in } Q^T,\\
&-p_1(t, \sigma) \textstyle\int_\Omega k_2(\sigma, \xi) y_2(t, \xi)\, d\xi \in \mu_2 \partial h_2(\omega_2(t, \sigma)) \quad \text{a.e. in } \Sigma_1^T.
\end{aligned} \tag{6.6'}$$

Proof. Let us put

$$H(t, y, \omega, p) = \langle F_1(t, y, \omega), p_1 \rangle + (F_2(t, y, \omega), p_2) + p_3 \int_\Omega f(y_2(x))\, dx$$

for all $t \in \mathbf{R}$, $y = (y_1, y_2, y_3) \in X^0 \times X^0 \times \mathbf{R}$, $p = (p_1, p_2, p_3) \in H \times X \times \mathbf{R}$, $\omega \in U$.

Evidently, B' is convex with nonempty interior. Hence by the preceding discussion, Corollary 2.5, and Remark 4.3, there exist $\lambda_0 \in \{0, 1\}$, μ_1, μ_2, $p_1 = p_1(t)$, $p_2 = p_2(t)$, $p_3 = p_3(t)$ such that

$$(l_0, \mu_1, \mu_2, p_2(T), p_3(T)) \neq 0 \tag{6.21}$$

and (p_1, p_2, p_3) satisfy the adjoint equations

$$\begin{aligned}
&p_1'(t) - A p_1(t) = -H_{y_1}(t, y(t), \omega(t), p(t)) \quad \text{a.e. in } [0, T],\\
&p_2'(t) = -H_{y_2}(t, y(t), \omega(t), p(t)) \quad \text{a.e. in } [0, T],\\
&p_3'(t) = -H_{y_3}(t, y(t), \omega(t), p(t)) = 0 \quad \text{a.e. in } [0, T],
\end{aligned}$$

the maximum principle

$$\begin{aligned}
&\int_0^T H(t, y(t), \omega(t), p(t)) + \mu_1 \int_\Omega h_1(\omega_1(t, x))\, dx + \mu_2 \int_{\Gamma_1} h_2(\omega_2(t, \sigma))\, d\sigma\, dt\\
&\quad \leqq \int_0^T H(t, y(t), \bar{\omega}(t), p(t)) + \mu_1 \int_\Omega h_1(\bar{\omega}_1(t, x))\, dx + \mu_2 \int_{\Gamma_1} h_2(\bar{\omega}_2(t, \sigma))\, d\sigma\, dt
\end{aligned}$$

for all $\bar{\omega} \in \mathcal{U}_{ad}(0, T)$, and the transversality conditions

$$
\begin{aligned}
&p_1(T) \equiv 0, \\
&(p_2(T), y_2 - y_2(T)) + p_3(T)(y_3 - y_3(T)) \leqq 0 \quad \text{for all } (y_2, y_3) \in B', \\
&\mu_i \geqq 0, \quad \mu_i = 0 \quad \text{if } \int_0^T \int h_i(\omega_i(t, x))\, dx\, dt < c_i \quad (i = 1, 2), \\
&H(T, y(T), \omega(T), p(T)) = -\lambda_0.
\end{aligned} \tag{6.22}
$$

Then by the same arguments used in the proof of Theorem 6.1, we have that (p_1, p_2) are solutions of the adjoint equations (6.3) with boundary conditions (6.4) and satisfy conditions (6.6′).

Finally, from (6.21) and (6.22) we deduce the remaining assertions, since by the second condition of (6.22) we have

$$p_1(T, x) \equiv 0,$$

$$p_2(T, x) \equiv \text{const.} = \mu_0 \geqq 0 \quad \text{and} \quad \mu_0 = 0 \quad \text{if } \int_\Omega y_2(T, x)\, dx < \bar{y}_2,$$

$$p_3(t) \equiv p_3(T) = \mu_3 \geqq 0 \quad \text{and} \quad \mu_3 = 0 \quad \text{if } y_3(T) = \int_{Q^T} f(y_2(t, x))\, dx\, dt < c_3.$$

Remark 6.3. If $\gamma = \gamma(t)$ is strictly increasing and $(T, y_1, y_2, \omega_1, \omega_2)$ is a solution of the optimal control problem: "Minimize the functional $J(T, x, u) = \gamma(T)$ among all the solutions of (6.1) satisfying (6.16) and (6.17)," then evidently the assertion of Theorem 6.2 still holds. By Remark 5.4 this is also true when γ is increasing and left differentiable.

7. Appendix. In this Appendix we will prove the following theorem.

THEOREM 7.1. *With the notation of* § 2, *for all* $t \in [0, T]$ *let* $B(t)$ *be a linear continuous mapping from* H *into* X_1 *such that* $B(t)x$ *is strongly measurable in* t *for all* x *and* $\|B(t)\| \leqq L(t)$ *for some* $L \in L^q(0, T; \mathbf{R})$, $q > 2$. *Then for every* $\xi \in X_1$, *the Cauchy problem*

$$
\begin{aligned}
&y'(t) + Ay(t) = B(t)y(t), \qquad t \in [0, T], \\
&y(0) = \xi
\end{aligned} \tag{7.1}
$$

has a unique solution y *in* $C(0, T; X_1) \cap L^2(0, T; H)$ *with* $y - G(\cdot)\xi \in W^{1,q^*}(0, T; X_1)$, $q^* = 2q/(2+q)$.

Actually, $y - G(\cdot)\xi \in W^{1,2}(0, T; X_1) \cap C(0, T; H)$ (*and therefore* $y \in W(0, T; H)$) *if* $q = +\infty$.

Moreover, if $\xi \in H$ *then we have* $y \in W^{1,2}(0, T; X_1) \cap C(0, T; H)$, *also when* $q = 2$.

To prove the theorem it is useful to premise the following lemma.

LEMMA 7.2. *Let us fix* $\alpha > 1$ *and for all* $g \in L^\alpha(0, T; X_1)$ *let us denote by* $S(g)$ *the unique solution of the Cauchy problem* $y' + Ay = g$ *in* $[0, T]$, $y(0) = 0$. *Then we have that*

$$S(g) \in W^{1,\alpha}(0, T; X_1); \tag{7.2}$$

$$S(g) \in L^\beta(0, T; H) \quad \text{for all } \beta < 2\alpha/(2-\alpha) \quad \text{if } \alpha < 2; \tag{7.3}$$

(7.4) $S(g) \in C(0, T; H)$ *and* $\|S(g)(t)\|_H^2 \leqq c \int_0^t |g(t)|^2\, dt$ *for all* $t \in [0, T]$ *if* $\alpha \geqq 2$.

Proof. The first part of the assertion follows from Theorem 4.1 of [9].

On the other hand, if $\alpha < 2$, then (for $\varepsilon < 1/2$ fixed), by [10, Thm. 19] we have that $y \in W^{\varepsilon,\alpha}(0, T; D_A(\theta - \varepsilon, \alpha))$ for all $\theta < 1$ such that $\theta - \varepsilon > 1/2$, and therefore $y \in W^{\varepsilon,\alpha}(0, T; D_A(1/2, 2)) = W^{\varepsilon,\alpha}(0, T; H)$ by [9, property (3.13), p. 325]. Hence (7.3) follows from the embedding of $W^{\varepsilon,\alpha}(0, T; H)$ into $L^{\alpha/(1-\varepsilon\alpha)}(0, T; H)$ and the fact that $\varepsilon < 1/2$ has been arbitrarily chosen.

Finally, (7.4) follows from Theorem II.1 and Lemma II.1 of [19].

Proof of Theorem 7.1. *Existence.* First, note that if $z \in L^\alpha(0, T; H)$, then $Bz = B(\cdot)z(\cdot) \in L^{\alpha q/(\alpha+q)}(0, T; X_1)$ and therefore by Lemma 7.2, we have

$$(7.5) \qquad S(Bz) \in \begin{cases} L^\beta(0, T; H) & \text{for all } \beta < \left(\frac{1}{\alpha} + \frac{1}{q} - \frac{1}{2}\right)^{-1} \quad \text{if } \alpha < \frac{2q}{q-2} \\ C(0, T; H) & \text{if } \alpha \geqq \frac{2q}{q-2}. \end{cases}$$

Now let us put

$$y_0(t) = G(t)\xi = z_0(t), \quad y_{n+1} = y_0 + S(By_n), \quad z_{n+1} = y_{n+1} - y_n = S(Bz_n).$$

Then by (7.5) it is easy to see that there exists ν such that $z_n \in C(0, T; H)$, for all $n \geqq \nu$. (Actually we have $\nu = 0$ if $\xi \in H$, $q \geqq 2$, and $\nu = 1$ if $q = +\infty$.) Moreover, from (7.4) we easily deduce by induction that for all $n \geqq \nu$ and for all $t \in [0, T]$ we have

$$\|z_n(t)\|_H^2 \leqq K \frac{(c \int_0^t L^2(s)\, ds)^{n-\nu}}{(n-\nu)!} \leqq K \frac{(c \int_0^T L^2(s)\, ds)^{n-\nu}}{(n-\nu)!},$$

where $K = \sup_{0 \leqq t \leqq T} \|z_\nu(t)\|_H^2$.

From this it follows that the series $\sum_{n=\nu}^\infty z_n$ converges in $C(0, T; H)$ to some z. Hence $y = z + z_0 + z_1 + \cdots + z_{\nu-1}$ belongs to $C(0, T; X_1) \cap L^2(0, T; H)$ and $y_n - y \to 0$ in $C(0, T; H)$ as $n \to \infty$. From this and from (7.4) it follows that $S(B(y_n - y)) \to 0$ in $C(0, T; H)$ and $y_n \to y$ in $C(0, T; X_1)$.

By the definition of y_n this proves that $y = y_0 + S(By)$, i.e., that y is a solution of (7.1); moreover, by (7.2) we have that $y - y_0 = S(By)$ belongs to $W^{1,q^*}(0, T; X_1)$, $q^* = 2q/(2+q)$, since $By \in L^{q^*}(0, T; X_1)$.

Finally, note that if $q = +\infty$ then we have $\nu = 1$; hence $y - y_0 = z \in C(0, T; H)$ and $y - y_0 = S(By) \in W^{1,2}(0, T; X_1)$ by (7.2). On the other hand, if $\xi \in H$ then $\nu = 0$ and therefore $y = z \in C(0, T; H)$ and $y = y_0 + S(By) \in W^{1,2}(0, T; X_1)$.

Uniqueness. Let y, z be solutions of (7.1); then we have that $y - z = S(B(y - z))$. From this by (7.5) and a bootstrap argument it follows that $y - z$ and $B(y - z)$ belong to $C(0, T; H)$. Therefore by (7.3) and the Gronwall inequality (see Remark 3.5), we easily have $y - z = 0$.

Acknowledgment. We thank V. Capasso for having supplied us with the manuscript of [1] as well as for some useful talks on the subject of this paper.

REFERENCES

[1] V. ARNAUTU, V. BARBU, AND V. CAPASSO, *Controlling the spread of a class of epidemics*, Appl. Math. Optim., 20 (1989), pp. 297–317.

[2] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Applications of Mathematics, 3, Springer-Verlag, New York, 1976.

[3] V. BARBU, *Optimal Control of Variational Inequalities*, Research Notes in Mathematics, 100, Pitman, London, 1984.

[4] V. BARBU AND T. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Sijthoff and Noordhoff, Bucuresti, 1978.

[5] N. BASILE AND M. MININNI, *A vector valued optimization approach to the study of a class of epidemics*, J. Math. Anal. Appl., to appear.

[6] L. D. BERKOVITZ, *Optimal Control Theory*, Applied Mathematical Sciences, 12, Springer-Verlag, New York, 1974.

[7] V. CAPASSO, *Asymptotic stability for an integrodifferential reaction diffusion system*, J. Math. Anal. Appl., 103 (1984), pp. 575–588.

[8] V. CAPASSO AND K. KUNISCH, *A reaction-diffusion system arising in modelling man-environment diseases*, Quart. Appl. Math., 46 (1988), pp. 431–450.

[9] G. DA PRATO AND P. GRISVARD, *Sommes d' opérateurs linéaires et équations différentielles operationélles*, J. Math. Pures Appl., 54 (1975), pp. 305–387.

[10] G. DI BLASIO, *Linear parabolic evolution equations in L^p-spaces*, Ann. Mat. Pura Appl., 138 (1984), pp. 55–104.

[11] J. V. EGOROV, *Optimal control in Banach spaces*, Soviet Math. Dokl., 4 (1963), pp. 630–633.

[12] H. O. FATTORINI, *A unified theory of necessary conditions for nonlinear nonconvex systems*, Appl. Math. Optim., 15 (1987), pp. 141–185.

[13] H. O. FATTORINI AND H. FRANKOWSKA, *Necessary conditions for infinite dimensional problems*, preprint.

[14] J. K. HALE, *Ordinary Differential Equations*, Pure and Applied Mathematics, Wiley-Interscience, New York, 1969.

[15] X. J. LI AND S. N. CHOW, *Maximum principle of optimal control for functional differential systems*, J. Optim. Theory Appl., 54 (1987), pp. 335–360.

[16] X. J. LI AND Y. L. YAO, *Maximum principle of distributed parameter systems with time lags*, in Proc. Conference on Control Theory of Distributed Parameters Systems and Applications, F. Kappel, K. Kunish, and W. Schappacher, eds., Lecture Notes in Control and Information Science, 75, Springer-Verlag, New York, 1985, pp. 410–427.

[17] J. L. LIONS, *Contrôle optimal des systemes gouvernés par des equations aux derivées partielles*, Dunod-Gauthier Villars, Paris, 1968.

[18] H. TANABE, *Equations of evolution*, Monographs and Studies in Mathematics, 6, Pitman, Boston, 1979.

[19] W. VON WAHL, *The equation $u' + A(t)u = f$ in a Hilbert space and L^p-estimates for parabolic equations*, J. London Math. Soc. (2), 25 (1982), pp. 483–497.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 1136-1147, September 1990

006

AN APPROXIMATION THEOREM FOR THE ALGEBRAIC RICCATI EQUATION*

FRANZ KAPPEL† AND DIETMAR SALAMON‡

Abstract. For an infinite-dimensional linear quadratic control problem in Hilbert space, approximation of the solution of the algebraic Riccati operator equation in the strong operator topology is considered under conditions weaker than uniform exponential stability of the approximating systems. As an application, strong convergence of the approximating Riccati operators in case of a previously developed spline approximation scheme for delay systems is established. Finally, convergence of the transfer-functions of the approximating systems is investigated.

Key words. linear quadratic control problem in Hilbert space, algebraic Riccati equation, hereditary control systems, spline approximation

AMS(MOS) subject classifications. 34K35, 41A15, 93D15

1. Introduction and hypotheses. Let H, U, and Y be Hilbert spaces, and consider the linear system

$$\begin{aligned} \dot{z}(t) &= Az(t)+Bu(t), \qquad z(0)=\varphi \in H, \\ y(t) &= Cz(t), \end{aligned} \tag{1.1}$$

where $A:\operatorname{dom} A \to H$ is the infinitesimal generator of a strongly continuous semigroup $S(t)\in \mathscr{L}(H)$, and $B\in \mathscr{L}(U, H)$, $C\in \mathscr{L}(H, Y)$ are bounded linear operators. Associated with (1.1) we consider the *algebraic Riccati equation*

$$\langle A\psi, P\varphi\rangle+\langle P\psi, A\varphi\rangle-\langle B^*P\psi, B^*P\varphi\rangle+\langle C\psi, C\varphi\rangle=0 \tag{1.2}$$

for φ, $\psi\in \operatorname{dom} A$. This equation has a nonnegative operator solution $P=P^*\in \mathscr{L}(H)$ if and only if for every $\varphi\in H$ there exists a control function $u\in L^2(0,\infty; U)$ such that the integral

$$J(u)=J(u,\varphi)=\int_0^\infty (\|u(t)\|^2+\|y(t)\|^2)\,dt \tag{1.3}$$

is finite. Under this assumption for every $\varphi\in H$ there exists a unique optimal control that is given by the feedback law

$$u(t)=-B^*Pz(t),$$

where P is the minimal nonnegative solution of (1.2). A nonnegative solution of (1.2) exists under the assumption that system (1.1) is *stabilizable*, meaning that there exists an operator $K\in \mathscr{L}(H, U)$ such that $A+BK$ generates an exponentially stable semigroup. If (1.1) is also *detectable* in the sense that for some operator $L\in \mathscr{L}(Y, H)$ the operator $A+LC$ generates an exponentially stable semigroup, then the solution P of (1.2) is unique in the class of nonnegative operators on H and the closed-loop semigroup generated by $A-BB^*P$ is exponentially stable [1], [6].

* Received by the editors December 2, 1988; accepted for publication (in revised form) October 30, 1988.

† Institut für Mathematik, Universität Graz, Elisabethstrasse 16, A 8010 Graz, Austria. The work of this author was supported by the FWF (Austria) under grant S3206.

‡ Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom. The work of this author was supported by Nuffield Foundation grant SCI/180/173/G.

Together with (1.1) we also consider a sequence of approximating control systems

$$\begin{aligned} \dot{z}^N(t) &= A^N z^N(t) + B^N u^N(t), \qquad z^N(0) = \pi^N \varphi, \\ y^N(t) &= C^N z^N(t), \end{aligned} \tag{1.4}$$

where $z^N \in \mathbf{R}^{k(N)}$, $u^N \in \mathbf{R}^{m(N)}$, $y^N \in \mathbf{R}^{p(N)}$, and A^N, B^N, C^N are matrices of suitable dimensions. We assume that there exist injective linear maps

$$\iota^N : \mathbf{R}^{k(N)} \to H, \quad j^N : \mathbf{R}^{m(N)} \to U, \quad k^N : \mathbf{R}^{p(N)} \to Y$$

and surjective linear maps

$$\pi^N : H \to \mathbf{R}^{k(N)}, \quad \rho^N : U \to \mathbf{R}^{m(N)}, \quad \sigma^N : Y \to \mathbf{R}^{p(N)}$$

such that $\pi^N \iota^N$, $\rho^N j^N$, $\sigma^N k^N$ are identity maps and $\iota^N \pi^N$, $j^N \rho^N$, $k^N \sigma^N$ are orthogonal projections. On the spaces $\mathbf{R}^{k(N)}$, $\mathbf{R}^{m(N)}$, and $\mathbf{R}^{p(N)}$ we will always consider the induced inner products $\langle z, w\rangle_N = \langle \iota^N z, \iota^N w\rangle_H$, $z, w \in \mathbf{R}^{k(N)}$, $\langle u, v\rangle_N = \langle j^N u, j^N v\rangle_U$, $u, v \in \mathbf{R}^{m(N)}$, and $\langle x, y\rangle_N = \langle k^N x, k^N y\rangle_Y$, $x, y \in \mathbf{R}^{p(N)}$. $(A^N)^*$, $(B^N)^*$, $(C^N)^*$, $\cdots$ always denote the adjoint matrices with respect to the induced inner products.

The purpose of this paper is to investigate the convergence properties of the solution matrices $P^N = (P^N)^*$ of the approximating algebraic Riccati equations

$$(A^N)^* P^N + P^N A^N - P^N B^N (B^N)^* P^N + (C^N)^* C^N = 0. \tag{1.5}$$

To formulate the results we introduce the following concepts. The approximating systems (1.4) are called *strongly convergent* to (1.1) if

$$S(t)\varphi = \lim_{N\to\infty} \iota^N e^{A^N t} \pi^N \varphi, \qquad S(t)^*\varphi = \lim_{N\to\infty} \iota^N e^{(A^N)^* t} \pi^N \varphi \tag{1.6}$$

uniformly on compact time intervals for all $\varphi \in H$,

$$\begin{aligned} &\iota^N B^N \rho^N \to B, \quad j^N (B^N)^* \pi^N \to B^*, \quad \text{and} \\ &j^N \rho^N \to \mathrm{id}_U \quad \text{strongly} \end{aligned} \tag{1.7}$$

and

$$\begin{aligned} &k^N C^N \pi^N \to C, \quad \iota^N (C^N)^* \sigma^N \to C^*, \quad \text{and} \\ &k^N \sigma^N \to \mathrm{id}_Y \quad \text{strongly.} \end{aligned} \tag{1.8}$$

We will call systems (1.4) *uniformly output stable* if there exists a constant $c > 0$ such

$$\int_0^\infty \| k^N C^N e^{A^N t} \pi^N \varphi \|^2 \, dt \leqq c \| \varphi \|^2$$

for all $\varphi \in H$ and $N = 1, 2, \cdots$. Systems (1.4) are said to be *uniformly input-output stable* if the functions $C^N e^{A^N t} B^N$, $N = 1, 2, \cdots$, are integrable on $0 \leqq t < \infty$ and there exists a constant $c_1 > 0$ such that

$$\| k^N C^N (i\omega I - A^N)^{-1} B^N \rho^N \| \leqq c_1$$

for all $\omega \in \mathbf{R}$ and $N = 1, 2, \cdots$.

Remarks. (1) Uniform output stability of systems (1.4) in connection with strong convergence to (1.1) implies that system (1.1) is output stable in the sense that

$$\int_0^\infty \| CS(t)\varphi \|^2 \, dt \leqq \text{const.} \, \| \varphi \|^2 \quad \text{for all } \varphi \in H.$$

(2) If the approximating systems (1.4) are strongly convergent to system (1.1) and the matrices $K^N \in \mathbf{R}^{m(N)\times k(N)}$, $L^N \in \mathbf{R}^{k(N)\times p(N)}$ are chosen such that the operator sequences $j^N K^N \pi^N \in \mathscr{L}(H, U)$, $\iota^N L^N \sigma^N \in \mathscr{L}(Y, H)$ and their adjoints $\iota^N (K^N)^* \rho^N$, $k^N (L^N)^* \pi^N$ converge strongly to K, L and K^*, L^*, respectively, then the feedback

systems

$$\begin{aligned} \dot{z}^N(t) &= (A^N + B^N K^N) z^N(t) + B^N v^N(t), \qquad z^N(0) = \pi^N \varphi, \\ y^N(t) &= C^N z^N(t), \qquad w^N(t) = K^N z^N(t), \end{aligned} \tag{1.9}$$

and the dynamic observers

$$\begin{aligned} \dot{z}^N(t) &= (A^N + L^N C^N) z^N(t) - L^N y^N(t) + B^N u^N(t), \\ z^N(0) &= \pi^N \varphi, \qquad w^N(t) = K^N z^N(t), \end{aligned} \tag{1.10}$$

are also strongly convergent. This can be seen by using the variation of parameters formula, Gronwall's inequality and Lebesgue's dominated convergence theorem.

(3) Let systems (1.4) converge strongly to system (1.1). By (1.6) and the uniform boundedness principle we see that there exists a constant $M_1 \geqq 1$ such that

$$\| \iota^N e^{A^N t} \pi^N \| \leqq M_1$$

for $t \in [0, 1]$ and $N = 1, 2, \cdots$. By standard considerations this implies

$$\| \iota^N e^{A^N t} \pi^N \| \leqq M_1 e^{\alpha t}, \quad t \geqq 0, \quad N = 1, 2, \cdots, \tag{1.11}$$

where α is some real constant. It follows that $\|S(t)\| \leqq M_1 e^{\alpha t}$, $t \geqq 0$. However, the exponential growth rate of $S(t)$ may be strictly less than $\alpha_0 = \inf \alpha$, where the infimum is over all α for which (1.11) holds with some constant $M_1 \geqq 1$. By the Trotter-Kato theorem we see that

$$\lim_{N \to \infty} \| (\lambda I - A)^{-1} z - \iota^N (\lambda I - A^N)^{-1} \pi^N z \| = 0$$

for all $z \in H$ uniformly for $\operatorname{Re} \lambda \geqq \gamma$ for any $\gamma > \alpha_0$.

(4) From the definition of the norms on $\mathbf{R}^{k(N)}$, $\mathbf{R}^{m(N)}$, and $\mathbf{R}^{p(N)}$ it is obvious that $\|\iota^N\| = \|j^N\| = \|k^N\| = 1$ and also $\|\pi^N\| = \|\rho^N\| = \|\sigma^N\| = 1$. Let $H^N = \operatorname{range} \iota^N \pi^N$, $U^N = \operatorname{range} j^N \rho^N$ and $Y^N = \operatorname{range} k^N \sigma^N$. Then, for instance,

$$\begin{aligned} \| \iota^N B^N \rho^N \|_{\mathscr{L}(U^N, H^N)} &= \| B^N \|_{\mathscr{L}(\mathbf{R}^{m(N)}, \mathbf{R}^{k(N)})}, \\ \| k^N C^N \pi^N \|_{\mathscr{L}(H^N, Y^N)} &= \| C^N \|_{\mathscr{L}(\mathbf{R}^{k(N)}, \mathbf{R}^{p(N)})}. \end{aligned}$$

In § 2 we will make use of these observations repeatedly.

2. The convergence result. The following theorem is the main result of this paper.

THEOREM 1. *Let systems* (1.4), $K^N \in \mathbf{R}^{m(N) \times k(N)}$, $L^N \in \mathbf{R}^{k(N) \times p(N)}$, $K \in \mathscr{L}(H, U)$, *and* $L \in \mathscr{L}(Y, H)$ *be given. Assume that*

(i) *Systems* (1.4) *are strongly convergent to* (1.1);

(ii) $j^N K^N \pi^N \to K$, $\iota^N (K^N)^* \rho^N \to K^*$, $\iota^N L^N \sigma^N \to L$, $k^N (L^N)^* \pi^N \to L^*$ *strongly*;

(iii) $A + BK$ *and* $A + LC$ *generate exponentially stable semigroups*;

(iv) *Systems* (1.9) *are uniformly output stable and uniformly input-output stable*; *and*

(v) *Systems* (1.10) *are uniformly input-output stable.*

Then

$$P\varphi = \lim_{N \to \infty} \iota^N P^N \pi^N \varphi$$

for every $\varphi \in H$, *where* $P \in \mathscr{L}(H)$ *and* $P^N \in \mathbf{R}^{k(N) \times k(N)}$ *are the minimal nonnegative solutions of* (1.2) *and* (1.5), *respectively.*

An earlier version of this convergence theorem was proved in [3] under stronger assumptions. In particular the following property of the approximation scheme is assumed (see [3, Conjecture 7.1]): If the semigroup $S(t)$ is exponentially stable, then the approximating semigroups satisfy an estimate $\|\exp(A^N t)\| \leqq M e^{-\beta t}$, $t \geqq 0$, $N =$

$1, 2, \cdots$, with constants $M \geqq 1$, $\beta > 0$ independent of N. This assumption is not met by the spline approximation scheme for delay systems developed in [4] and [5]. On the other hand, in this case convergence of the P^N's has been observed numerically [4]. In § 3 we will show that the spline scheme indeed satisfies the requirements of Theorem 1.

The proof of Theorem 1 rests on the relationship between the algebraic Riccati equation (1.2) and the optimal control problem (1.3). We first establish two lemmas. For system (1.1), respectively systems (1.4), we define the operators $\mathscr{E}$, $\mathscr{E}^N : H \to L^2(0, \infty; Y)$ by

$$(\mathscr{E}\varphi)(t) = CS(t)\varphi, \qquad t \geqq 0,$$
$$(\mathscr{E}^N\varphi)(t) = k^N C^N e^{A^N t}\pi^N\varphi, \qquad t \geqq 0,$$

respectively. Then the adjoint operators $\mathscr{E}^*$, $(\mathscr{E}^N)^* : L^2(0, \infty; Y) \to H$ are given by

$$\mathscr{E}^* y = \int_0^\infty S(t)^* C^* y(t)\, dt,$$

and

$$(\mathscr{E}^N)^* y = \int_0^\infty \iota^N e^{(A^N)^* t}(C^N)^* \sigma^N y(t)\, dt.$$

LEMMA 1. *Assume that $S(t)$ is exponentially stable and systems* (1.4) *are uniformly output stable and converge strongly to system* (1.1). *Then*

$$\mathscr{E}^N \to \mathscr{E} \quad \textit{and} \quad (\mathscr{E}^N)^* \to \mathscr{E}^*$$

strongly as $N \to \infty$.

Proof. For any $T > 0$ we get

$$\begin{aligned}\|\mathscr{E}\varphi - \mathscr{E}^N\varphi\|^2_{L^2(0,\infty;Y)} \leqq \int_0^T \|CS(t)\varphi - k^N C^N e^{A^N t}\pi^N\varphi\|^2\, dt + 3\int_0^\infty \|CS(T+t)\varphi\|^2\, dt \\ + 3\int_0^\infty \|k^N C^N e^{A^N t}(e^{A^N T}\pi^N\varphi - \pi^N S(T)\varphi)\|^2\, dt \\ + 3\int_0^\infty \|k^N C^N e^{A^N t}\pi^N S(T)\varphi\|^2\, dt =: \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4.\end{aligned}$$

The estimate for α_1 is

$$\begin{aligned}\alpha_1 \leqq 2\int_0^T \|(C - k^N C^N \pi^N)S(t)\varphi\|^2\, dt \\ + 2\sup_N \|k^N C^N \pi^N\|^2 \int_0^T \|S(t)\varphi - \iota^N e^{A^N t}\pi^N\varphi\|^2\, dt.\end{aligned}$$

For any $T > 0$ the right-hand side tends to zero as $N \to \infty$, because systems (1.4) are strongly convergent to (1.1).

For α_2 we get from the exponential stability of $S(t)$ (i.e., $\|S(t)\| \leqq Me^{-\beta t}$, $t \geqq 0$, for some $\beta > 0$)

$$\alpha_2 \leqq 3\|C\|^2 e^{-2\beta T}\frac{M^2}{2\beta}\|\varphi\|^2.$$

Using uniform output stability of systems (1.4) we obtain

$$\alpha_3 \leqq 3c\|\iota^N e^{A^N T}\pi^N\varphi - S(T)\varphi\|^2,$$
$$\alpha_4 \leqq 3c\|S(T)\varphi\|^2 \leqq 3cM^2 e^{-2\beta T}\|\varphi\|^2.$$

These estimates together show that

$$\mathscr{E}^N\varphi \to \mathscr{E}\varphi \quad \text{as } N\to\infty$$

for any $\varphi \in H$.

For the proof of $(\mathscr{E}^N)^*y \to \mathscr{E}^*y$ it is enough to consider y with compact support. Let supp $y \subset [0, T]$, $T>0$. Then

$$\|\mathscr{E}^*y-(\mathscr{E}^N)^*y\| \leqq \int_0^T \|S(t)^*C^*y(t)-\iota^N e^{(A^N)^*t}\pi^N C^*y(t)\|\,dt$$
$$+\int_0^T \|\iota^N e^{(A^N)^*t}\pi^N\|\,\|(C^*-\iota^N(C^N)^*\sigma^N)y(t)\|\,dt.$$

The right-hand side tends to zero by the Lebesgue dominated convergence theorem using strong convergence of systems (1.4) to (1.1) and (1.11). □

Remark. If dim $Y<\infty$ and systems (1.4) are uniformly exponentially stable (i.e., $\|\iota^N e^{A^Nt}\pi^N\| \leqq Me^{-\alpha t}$, $t\geqq 0$, $N=1, 2, \cdots$, for some constants $M\geqq 1$, $\alpha>0$), then

$$\|\mathscr{E}-\mathscr{E}^N\| = \|\mathscr{E}^*-(\mathscr{E}^N)^*\| \to 0 \quad \text{as } N\to\infty.$$

We only have to observe that $\|CS(t)-k^NC^Ne^{A^Nt}\pi^N\|$ is exponentially decaying as $t\to\infty$ uniformly with respect to N.

To state the next lemma we introduce the operators $\mathscr{G}$, $\mathscr{G}^N: L^2(0,\infty; U)\to L^2(0,\infty; Y)$ by

$$(\mathscr{G}u)(t)=\int_0^t CS(t-s)Bu(s)\,ds, \qquad t\geqq 0,$$

$$(\mathscr{G}^Nu)(t)=\int_0^t k^NC^Ne^{A^N(t-s)}B^N\rho^Nu(s)\,ds, \qquad t\geqq 0,$$

for $u\in L^2(0,\infty; U)$. The adjoint operators are given by

$$(\mathscr{G}^*y)(t)=\int_t^\infty B^*S(s-t)^*C^*y(s)\,ds, \qquad t\geqq 0,$$

$$((\mathscr{G}^N)^*y)(t)=\int_t^\infty j^N(B^N)^*e^{(A^N)^*(s-t)}(C^N)^*\sigma^Ny(s)\,ds, \qquad t\geqq 0,$$

for $y\in L^2(0,\infty; Y)$.

LEMMA 2. *Assume that $S(t)$ is exponentially stable, systems* (1.4) *are strongly convergent to system* (1.1) *and, furthermore, systems* (1.4) *are uniformly input-output stable and uniformly output stable. Then*

$$\mathscr{G}^N\to\mathscr{G} \quad \textit{and} \quad (\mathscr{G}^N)^*\to\mathscr{G}^*$$

strongly as $N\to\infty$.

Proof. Using Parseval's equality and uniform input-output stability we obtain the estimate

$$\|\mathscr{G}^Nu\|^2=\int_0^\infty \left\|\int_0^t k^NC^Ne^{A^N(t-s)}B^N\rho^Nu(s)\,ds\right\|^2 dt$$
$$\leqq \int_{-\infty}^\infty \|k^NC^N(i\omega I-A^N)^{-1}B^N\rho^N\|^2\|\hat{u}(\omega)\|^2\,d\omega$$
$$\leqq c_1^2\int_{-\infty}^\infty \|\hat{u}(\omega)\|^2\,d\omega = c_1^2\|u\|^2,$$

which implies uniform boundedness of the operators $\mathcal{G}^N$. Therefore it is enough to consider input functions u with compact support, $\operatorname{supp} u \subset [0, T]$. Let $y = \mathcal{G}u$ and $y^N = \mathcal{G}^N u$. Then the estimate

$$\begin{aligned}\|y(t)-y^N(t)\| \leqq & \int_0^t \|(C-k^N C^N \pi^N) S(s) B u(t-s)\|\, ds \\ & + \sup_N \|k^N C^N \pi^N\| \int_0^t \|(S(s)-\iota^N e^{A^N s}\pi^N) B u(t-s)\|\, ds \\ & + M_1 e^{\alpha T} \sup_N \|k^N C^N \pi^N\| \int_0^t \|(B-\iota^N B^N \rho^N) u(t-s)\|\, ds\end{aligned}$$

for $0 \leqq t \leqq T$, shows that $\|y(\cdot)-y^N(\cdot)\|_{L^2(0,T;Y)} \to 0$ as $N \to \infty$ (using strong convergence of systems (1.4) to system (1.1) and the Lebesgue dominated convergence theorem). Moreover, we have

$$y(t+T) = (\mathcal{E}\varphi)(t) \quad \text{with } \varphi = \int_0^T S(T-s) B u(s)\, ds,$$

$$y^N(t+T) = (\mathcal{E}^N \varphi^N)(t) \quad \text{with } \varphi^N = \iota^N \int_0^T e^{A^N(T-s)} B^N \rho^N u(s)\, ds$$

and hence it follows from Lemma 1 that $y^N \to y$ in $L^2(T, \infty; Y)$ as $N \to \infty$.

For the adjoint operators we again need to consider y with compact support only, say $\operatorname{supp} y \subset [0, T]$, $T > 0$. Then

$$\|(\mathcal{G}^N)^* y - \mathcal{G}^* y\|^2 = \int_0^T \left\| \int_t^T (B^* S(s-t)^* C^* - j^N (B^N)^* e^{(A^N)^*(s-t)} (C^N)^* \sigma^N) y(s)\, ds \right\|^2 dt.$$

Using strong convergence of systems (1.4) to (1.1) (together with the estimate (1.11)) we see that we can apply the Lebesgue dominated convergence theorem twice. □

Proof of Theorem 1. Let $S_K(t)$ and $S_L(t)$ denote the semigroups generated by $A+BK$ and $A+LC$, respectively. We first observe that $J(u) < \infty$ for $u \in L^2(0, \infty; U)$ if and only if $v = u - Kz \in L^2(0, \infty; U)$, where $z(t)$ is the mild solution of $\dot{z}(t) = Az(t) + Bu(t)$, $z(0) = \varphi$, i.e., $z(t) = S(t)\varphi + \int_0^t S(t-s) B u(s)\, ds$. Indeed, since

$$z(t) = S_K(t)\varphi + \int_0^t S_K(t-s) B v(s)\, ds$$

(this is rather obvious for $\varphi \in \operatorname{dom} A$ and u being differentiable and follows by a density argument in the general case), $z(t)$ is square integrable if v is. But then $u = v + Kz \in L^2(0, \infty; U)$ and $y = Cz \in L^2(0, \infty; Y)$, i.e., $J(u) < \infty$. Conversely, if $J(u) < \infty$ then the formula

$$z(t) = S_L(t)\varphi + \int_0^t S_L(t-s)(B u(s) - L y(s))\, ds$$

shows that z and $v = u - Kz$ are square integrable.

Therefore the control problem of minimizing (1.3) subject to (1.1) is equivalent to the problem of minimizing

$$J_K(v) = J_K(v, \varphi) = \int_0^\infty (\|v(t) + Kz(t)\|^2 + \|y(t)\|^2)\, dt \tag{2.1}$$

subject to

$$\dot{z} = (A+BK)z + Bv, \quad z(0) = \varphi, \quad y = Cz. \tag{2.2}$$

The functional (2.1) is bounded for all $v \in L^2(0, \infty; U)$ and can be written in the form

$$J_K(v, \varphi) = \|\mathscr{C}\varphi + \mathscr{T}v\|^2,$$

where the operators $\mathscr{C}: H \to L^2(0, \infty; U \times Y)$ and $\mathscr{T}: L^2(0, \infty; U) \to L^2(0, \infty; U \times Y)$ are defined by

$$(\mathscr{C}\varphi)(t) = (KS_K(t)\varphi, CS_K(t)\varphi),$$

$$(\mathscr{T}v)(t) = \left(v(t) + K\int_0^t S_K(t-s)Bv(s)\,ds,\ C\int_0^t S_K(t-s)Bv(s)\,ds\right).$$

Hence the optimal control $\hat{v}$ satisfies

$$\mathscr{T}^*\mathscr{T}\hat{v} + \mathscr{T}^*\mathscr{C}\varphi = 0. \tag{2.3}$$

We define the operator $\mathscr{F}: L^2(0, \infty; U \times Y) \to L^2(0, \infty; U)$ by

$$\mathscr{F}(u, y)(t) = u(t) - K\int_0^t S_L(t-s)(Bu(s) - Ly(s))\,ds.$$

Then straightforward computations show that, for $t \geqq 0$,

$$(\mathscr{F}\mathscr{T}v)(t) = v(t) + Kz(t) - K\int_0^t S_L(t-s)(Bv(s) + BKz(s) - LCz(s))\,ds,$$

where $z(t) = \int_0^t S_K(t-s)Bu(s)\,ds$. Let $w(t)$ denote the integral term in the above equation. Then $w(t)$ is the unique mild solution of $\dot{w} = (A+LC)w + Bv(t) + BKz(t) - LCz(t)$, $w(0) = 0$. Obviously, $z(t)$ is also a mild solution of this problem, i.e., $w(t) \equiv z(t)$. Thus we have

$$(\mathscr{F}\mathscr{T}v)(t) = v(t) + Kz(t) - Kz(t) = v(t), \qquad t \geqq 0,$$

i.e.,

$$\mathscr{F}\mathscr{T}v = v \quad \text{for all } v \in L^2(0, \infty; U).$$

This implies $\|v\|^2 \leqq \|\mathscr{F}\|^2\|\mathscr{T}v\|^2 = \|\mathscr{F}\|^2\langle v, \mathscr{T}^*\mathscr{T}v\rangle \leqq \|\mathscr{F}\|^2\|v\|\,\|\mathscr{T}^*\mathscr{T}v\|$, i.e.,

$$\|\mathscr{T}^*\mathscr{T}v\| \geqq \|\mathscr{F}\|^{-2}\|v\| \quad \text{for all } v \in L^2(0, \infty; U).$$

Hence the operator $\mathscr{T}^*\mathscr{T}$ is boundedly invertible, and from (2.3) we get

$$\hat{v} = -(\mathscr{T}^*\mathscr{T})^{-1}\mathscr{T}^*\mathscr{C}\varphi.$$

The identity $J_K(\hat{v}, \varphi) = J(\hat{u}, \varphi) = \langle \varphi, P\varphi\rangle$ shows that

$$\langle \varphi, P\varphi\rangle = \langle \mathscr{C}\varphi, \mathscr{C}\varphi + \mathscr{T}\hat{v}\rangle = \langle \varphi, \mathscr{C}^*\mathscr{C}\varphi - \mathscr{C}^*\mathscr{T}(\mathscr{T}^*\mathscr{T})^{-1}\mathscr{T}^*\mathscr{C}\varphi\rangle$$

and hence

$$P = \mathscr{C}^*(I - \mathscr{T}(\mathscr{T}^*\mathscr{T})^{-1}\mathscr{T}^*)\mathscr{C}. \tag{2.4}$$

Defining the approximating operators

$$\mathscr{C}^N: \mathbf{R}^{k(N)} \to L^2(0, \infty; \mathbf{R}^{k(N)} \times \mathbf{R}^{p(N)}),$$

$$\mathscr{T}^N: L^2(0, \infty; \mathbf{R}^{m(N)}) \to L^2(0, \infty; \mathbf{R}^{m(N)} \times \mathbf{R}^{p(N)}),$$

$$\mathscr{F}^N: L^2(0, \infty; \mathbf{R}^{m(N)} \times \mathbf{R}^{p(N)}) \to L^2(0, \infty; \mathbf{R}^{m(N)})$$

in the obvious way we get analogously as above

$$P^N = (\mathscr{C}^N)^*(I - \mathscr{T}^N((\mathscr{T}^N)^*\mathscr{T}^N)^{-1}(\mathscr{T}^N)^*)\mathscr{C}^N. \tag{2.5}$$

Lemma 1 applied to systems (1.9) shows that

$$(j^N \oplus k^N)\mathscr{C}^N\pi^N \to \mathscr{C} \quad \text{and} \quad \iota^N(\mathscr{C}^N)^*(\rho^N \oplus \sigma^N) \to \mathscr{C}^* \quad \text{strongly.} \tag{2.6}$$

Here $j^N \oplus k^N$ denotes the direct sum of j^N and k^N defined by $(j^N \oplus k^N)(u^N, y^N) = (j^N u^N, k^N y^N)$, $u^N \in \mathbf{R}^{m(N)}$, $y^N \in \mathbf{R}^{p(N)}$ etc. Moreover, in abuse of notation we define $j^N u$ for $u \in L^2(0, \infty; \mathbf{R}^{m(N)})$ by $(j^N u)(t) = j^N u(t)$, $t \geqq 0$, etc.

By Lemma 2 applied to systems (1.9) we obtain

$$(j^N \oplus k^N)\mathscr{T}^N\rho^N \to \mathscr{T} \quad \text{and} \quad j^N(\mathscr{T}^N)^*(\rho^N \oplus \sigma^N) \to \mathscr{T}^* \quad \text{strongly.} \tag{2.7}$$

By assumption (v) of Theorem 1 we have

$$\sup_N \|\mathscr{F}^N\| < \infty.$$

This and the estimate

$$\|(\mathscr{T}^N)^*\mathscr{T}^N v^N\| \geqq \|\mathscr{F}^N\|^{-2}\|v^N\|$$

for all $v^N \in L^2(0, \infty; \mathbf{R}^{m(N)})$ show that $\|((\mathscr{T}^N)^*\mathscr{T}^N)^{-1}\|$ are uniformly bounded. By Remark (4) of § 1 also $\|j^N((\mathscr{T}^N)^*\mathscr{T}^N)^{-1}\rho^N\|$ are uniformly bounded. Then for $v \in L^2(0, \infty; U)$

$$\begin{aligned} j^N((\mathscr{T}^N)^*\mathscr{T}^N)^{-1}\rho^N v - (\mathscr{T}^*\mathscr{T})^{-1}v &= j^N((\mathscr{T}^N)^*\mathscr{T}^N)^{-1}\rho^N v - j^N\rho^N(\mathscr{T}^*\mathscr{T})^{-1}v \\ &\quad - ((\mathscr{T}^*\mathscr{T})^{-1}v - j^N\rho^N(\mathscr{T}^*\mathscr{T})^{-1}v). \end{aligned}$$

The second term on the right-hand side converges to zero as $N \to \infty$. For the first term we get

$$\begin{aligned} &j^N((\mathscr{T}^N)^*\mathscr{T}^N)^{-1}\rho^N v - j^N\rho^N(\mathscr{T}^*\mathscr{T})^{-1}v \\ &\quad = j^N((\mathscr{T}^N)^*\mathscr{T}^N)^{-1}\rho^N\big(\mathscr{T}^*\mathscr{T} - j^N(\mathscr{T}^N)^*\mathscr{T}^N\rho^N\big)(\mathscr{T}^*\mathscr{T})^{-1}v, \end{aligned}$$

which proves

$$j^N((\mathscr{T}^N)^*\mathscr{T}^N)^{-1}\rho^N \to (\mathscr{T}^*\mathscr{T})^{-1} \quad \text{strongly.} \tag{2.8}$$

The representations (2.4) and (2.5) together with (2.6)-(2.8) prove that

$$\iota^N P^N \pi^N \to P \quad \text{strongly.} \qquad \square$$

Remark. If the matrices $e^{(A^N + B^N K^N)t}$, $t \geqq 0$, are uniformly exponentially stable, then the operators $\mathscr{C}^N\pi^N$ converge to $\mathscr{C}$ in the uniform operator topology (see the remark following Lemma 1). Then it follows that the operators $\iota^N P^N \pi^N$ also converge in the uniform operator topology provided $\dim U < \infty$ and $\dim Y < \infty$. It remains an open question whether convergence of the P^N in the uniform operator topology can be established under weaker assumptions.

3. Spline approximation for delay equations. The system

$$\begin{aligned} &\dot{x}(t) = A_0x(t) + A_1x(t-h) + B_0u(t), \qquad y(t) = C_0x(t), \\ &x(0) = \varphi^0, \quad x(\tau) = \varphi^1(\tau) \quad \text{for } -h \leqq \tau < 0, \end{aligned} \tag{3.1}$$

with $x(t) \in \mathbf{R}^n$, $u(t) \in \mathbf{R}^m$, $y(t) \in \mathbf{R}^p$ and $\varphi = (\varphi^0, \varphi^1) \in \mathbf{R}^n \times L^2(-h, 0; \mathbf{R}^n)$ is equivalent to system (1.1) in the Hilbert space

$$H = M^2 = \mathbf{R}^n \times L^2(-h, 0; \mathbf{R}^n).$$

Operators A, B, and C are given by

$$\begin{aligned} &\operatorname{dom} A = \{\varphi = (\varphi^0, \varphi^1) \in M^2 \mid \varphi^1 \in W^{1,2}(-h, 0; \mathbf{R}^n),\ \varphi^0 = \varphi^1(0)\}, \\ &A = (A_0\varphi^0 + A_1\varphi^1(-h), \dot{\varphi}^1) \quad \text{for } \varphi \in \operatorname{dom} A, \\ &Bu = (B_0u, 0) \quad \text{for } u \in \mathbf{R}^m, \\ &C\varphi = C_0\varphi^0 \quad \text{for } \varphi \in M^2. \end{aligned}$$

In [4] and [5] we have considered a sequence of approximating systems (1.4) where $k(N) = n(N+2)$, $m(N) = m$, $p(N) = p$ and the matrices A^N, B^N, C^N are given by $A^N = (Q^N)^{-1}H^N$ with

$$Q^N = \begin{bmatrix} I & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \frac{h}{3N}I & \frac{h}{6N}I & 0 & \cdot & \cdot & 0 \\ 0 & \frac{h}{6N}I & \frac{2h}{3N}I & \cdot & \cdot & & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & & \cdot & \cdot & \frac{2h}{3N}I & \frac{h}{6N}I \\ 0 & 0 & \cdot & \cdot & 0 & \frac{h}{6N}I & \frac{h}{3N}I \end{bmatrix},$$

$$H^N = \begin{bmatrix} A_0 & 0 & \cdot & \cdot & \cdot & \cdot & A_1 \\ I & -\frac{1}{2}I & -\frac{1}{2}I & 0 & \cdot & \cdot & 0 \\ 0 & \frac{1}{2}I & 0 & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & & & \cdot & \cdot & 0 & -\frac{1}{2}I \\ 0 & \cdot & \cdot & \cdot & 0 & \frac{1}{2}I & -\frac{1}{2}I \end{bmatrix},$$

$$B^N = \begin{bmatrix} B_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad C^N = (C_0 \cdots 0).$$

The injections ι^N are given by

$$\iota^N z = \left(z_0, \sum_{j=0}^{N} z_{ij} s_j^N\right),$$

where $z = \operatorname{col}(z_0, z_{10}, \cdots, z_{1N}) \in \mathbf{R}^{n(N+2)}$ and the functions s_j^N are the basis splines

$$s_0^N(\tau) = \max\left(0, N\frac{\tau}{h} + 1\right), \qquad s_N^N(\tau) = \max\left(0, 1 - N - N\frac{\tau}{h}\right),$$

and, for $j=1,\cdots,N-1$,

$$s_j^N(\tau)=\begin{cases} N\frac{\tau}{h}+j+1 & \text{for } -(j+1)\frac{h}{N}\leqq\tau\leqq -j\frac{h}{N},\\ -N\frac{\tau}{h}-j+1 & \text{for } -j\frac{h}{N}\leqq\tau\leqq -(j-1)\frac{h}{N},\\ 0 & \text{elsewhere.}\end{cases}$$

The induced inner product on $\mathbf{R}^{n(N+2)}$ is given by $\langle z, w\rangle_N = z^TQ^Nw$. Of course, $U=U^N=\mathbf{R}^m$ and $Y=Y^N=\mathbf{R}^p$ for all N.

The approximating systems (1.4) with these matrices are strongly convergent [4] and if the delay system (3.1) is stable in the sense that $\operatorname{Re}\lambda<0$ for all roots of $\det(\lambda I-A_0-e^{-\lambda h}A_1)=0$, then the approximating systems (1.4) are uniformly output stable [5]. Moreover, the approximating transfer functions are in this case given by

$$C^N(i\omega I-A^N)^{-1}B^N = C_0(i\omega I-A_0-\alpha^N(i\omega)A_1)^{-1}B_0, \tag{3.2}$$

where $\alpha^N(\lambda)$ is a sequence of rational functions converging to $e^{-\lambda h}$ uniformly on compact sets and satisfying $|\alpha^N(\lambda)|\leqq 2$ on $\operatorname{Re}\lambda\geqq 0$ for all $N=1,2,\cdots$ [5]. This shows that the approximating systems (1.4) are uniformly input-output stable for N sufficiently large provided that the delay system (3.1) is stable (which is equivalent to exponential stability of the corresponding system (1.1)).

THEOREM 2. *Suppose that there exist matrices $K_0\in\mathbf{R}^{m\times n}$ and $L_0\in\mathbf{R}^{n\times p}$ such that the delay systems*

$$\dot{x}(t)=(A_0+B_0K_0)x(t)+A_1x(t-h),$$
$$\dot{x}(t)=(A_0+L_0C_0)x(t)+A_1x(t-h)$$

are stable and let the matrices A^N, B^N, C^N be defined as above. Then there exist unique nonnegative solutions

$$P\in\mathscr{L}(M^2) \quad \text{and} \quad P^N\in\mathbf{R}^{n(N+2)\times n(N+2)}$$

of (1.2) *and* (1.5), *respectively, and for every $\varphi\in M^2$*

$$P\varphi=\lim_{N\to\infty}\iota^NP^N\pi^N\varphi.$$

Proof. Define the matrices

$$K^N=(K_0 \quad 0 \quad \cdots \quad 0), \qquad L^N=\begin{pmatrix}L_0\\0\\\vdots\\0\end{pmatrix}$$

and apply Theorem 1. □

Numerical examples for this convergence result are reported in [4].

Remark. The conditions of Theorem 2 are stronger than stabilizability and detectability of the delay system (3.1). However, we are not aware of a stabilizable delay system that cannot be stabilized by a feedback law of the form $u(t)=K_0x(t)$.

4. Convergence of transfer functions. In this section we give a short discussion of the connection between strong convergence of systems (1.4) and convergence of the corresponding transfer functions on the imaginary axis.

If the semigroup $S(t)$ is exponentially stable and systems (1.4) are strongly convergent to system (1.1) and are uniformly input-output stable, then we can show that

$$\lim_{N\to\infty} k^N C^N(\lambda I - A^N)^{-1} B^N \rho^N = C(\lambda I - A)^{-1} B$$

uniformly on compact subsets of $\operatorname{Re}\lambda > 0$. The proof involves Vitali's theorem on sequences of holomorphic functions (see, for instance, [2, p. 309]). Despite the fact that under the assumption of uniform input-output stability the functions $k^N C^N(\lambda I - A^N)^{-1} B^N \rho^N$ are uniformly bounded on $\operatorname{Re}\lambda \geqq 0$ (and not only on compact subsets of $\operatorname{Re}\lambda > 0$ as required in Vitali's theorem) we cannot conclude uniform convergence of these functions on compact subsets of the imaginary axis. This is demonstrated by the following example.

Example. Let $H = l^2$ and $U = Y = \mathbf{R}$. For an element $b = (b_1, b_2, \cdots) \in l^2$ with $b_j > 0$ for all j we consider

$$\dot{z}(t) = -z(t) + bu(t), \quad t \geqq 0, \qquad y(t) = \langle b, z(t)\rangle_{l^2}. \tag{4.1}$$

The solution semigroup of the homogeneous problem is $S(t) = e^{-t} I$, which obviously is exponentially stable. We consider the approximating systems

$$\dot{z}^N(t) = A^N z^N(t) + b^N u(t), \quad t \geqq 0, \qquad y(t) = (b^N)^T z^N(t), \tag{4.2}$$

where

$$A^N = \operatorname{diag}(-1, \cdots, -1, -b_{N+1}^2) \in \mathbf{R}^{(N+1)\times(N+1)}$$

$$b^N = \operatorname{col}(b_1, \cdots, b_{N+1}) \in \mathbf{R}^{N+1}.$$

The embedding $\iota^N : \mathbf{R}^{N+1} \to l^2$ is given by $\iota^N z^N = (z_1, \cdots, z_{N+1}, 0, \cdots)$ for $z^N = \operatorname{col}(z_1, \cdots, z_{N+1}) \in \mathbf{R}^{N+1}$ and the "projections" π^N by $\pi^N z = \operatorname{col}(z_1, \cdots, z_{N+1})$ for $z = (z_1, z_2, \cdots) \in l^2$.

The solutions of $\dot{z} = -z$, $z(0) = \varphi = (\varphi_1, \varphi_2, \cdots) \in l^2$, and $\dot{z}^N = A^N z^N$, $z^N(0) = \pi^N \varphi$, are given by

$$z(t) = e^{-t}\varphi, \qquad z^N(t) = e^{-t} \sum_{j=1}^{N} \varphi_j + e^{-b_{N+1}^2 t} \varphi_{N+1},$$

respectively. Therefore

$$\|z(t) - \iota^N z^N(t)\|_{l^2}^2 = (e^{-t} - e^{-b_{N+1}^2 t})^2 |\varphi_{N+1}|^2 + e^{-2t} \sum_{j=N+2}^{\infty} |\varphi_j|^2$$

$$\leqq \sum_{N+1}^{\infty} |\varphi_j|^2 \to 0 \quad \text{as } N \to \infty.$$

It is obvious that $\|b - \iota^N b^N\|_{l^2} \to 0$ as $N \to \infty$. Thus systems (4.2) are strongly convergent to (4.1).

Obviously $(b^N)^T e^{A^N t} b^N$ is integrable on $t \geqq 0$. The transfer functions $G^N(\lambda) = (b^N)^T(\lambda I - A^N)^{-1} b^N$ are given by

$$G^N(\lambda) = \frac{1}{1+\lambda} \sum_{j=1}^{N} b_j^2 + \frac{b_{N+1}^2}{b_{N+1}^2 + \lambda}.$$

Therefore

$$|G^N(i\omega)| \leqq \frac{1}{|1+i\omega|} \|b\|_{l^2}^2 + \frac{b_{N+1}^2}{|b_{N+1}^2 + i\omega|} \leqq \|b\|_{l^2}^2 + 1$$

for all $\omega \in \mathbf{R}$ and $N = 1, 2, \cdots$, i.e., systems (4.2) are uniformly input-output stable.

For $\varphi = (\varphi_1, \varphi_2, \cdots) \in l^2$ we get

$$(b^N)^T e^{A^N t} \pi^N \varphi = e^{-t} \sum_{j=1}^{N} b_j \varphi_j + e^{-b_{N+1}^2 t} b_{N+1} \varphi_{N+1}.$$

Therefore

$$\begin{aligned} |(b^N)^T e^{A^N t} \pi^N \varphi|^2 &\leqq (e^{-t} \|b\|_{l^2} \|\varphi\|_{l^2} + e^{-b_{N+1}^2 t} b_{N+1} |\varphi_{N+1}|)^2 \\ &\leqq 2e^{-2t} \|b\|_{l^2}^2 \|\varphi\|_{l^2}^2 + 2e^{-2b_{N+1}^2 t} b_{N+1}^2 |\varphi_{N+1}|^2 \end{aligned}$$

and

$$\int_0^\infty |(b^N)^T e^{A^N t} \pi^N \varphi|^2 \, dt \leqq \|b\|_{l^2}^2 \|\varphi\|_{l^2}^2 + |\varphi_{N+1}|^2 \leqq (\|b\|_{l^2}^2 + 1) \|\varphi\|_{l^2}^2,$$

which proves uniform output stability of systems (4.2).

Finally, if we define

$$G(\lambda) = \langle b, (\lambda I - A)^{-1} b \rangle_{l^2} = \frac{1}{1+\lambda} \|b\|_{l^2}^2,$$

then we immediately see that for $\lambda \neq -1$ (note that $b_{N+1}^2 \to 0$ as $N \to \infty$)

$$\lim_{N \to \infty} G^N(\lambda) = \begin{cases} G(\lambda) & \text{for } \lambda \neq 0, \\ G(\lambda) + 1 & \text{for } \lambda = 0. \end{cases}$$

This example shows that even under additional assumptions we cannot obtain uniform convergence of the transfer functions of the approximating systems on bounded subsets of $\mathbf{R}$ in general. But we can prove the following proposition.

PROPOSITION 1. *Under the assumptions of Lemma* 1 *we have*

$$\int_{-\infty}^{\infty} \|C(i\omega I - A)^{-1} B\xi - k^N C^N (i\omega I - A^N)^{-1} B^N \rho^N \xi\|_Y^2 \to 0$$

as $N \to \infty$ *for any* $\xi \in U$.

Proof. Using Parseval's identity we get

$$\begin{aligned} &\int_{-\infty}^{\infty} \|C(i\omega I - A)^{-1} \varphi - k^N C^N (i\omega I - A^N)^{-1} \pi^N \varphi^N\|^2 \, d\omega \\ &\quad = \int_0^\infty \|CS(t)\varphi - k^N C^N e^{A^N t} \pi^N \varphi^N\|^2 \, dt = \|\mathscr{C}\varphi - \mathscr{C}^N \varphi^N\|_{L^2(0,\infty;Y)}^2 \end{aligned}$$

for φ, $\varphi^N \in H$. Hence the result follows from Lemma 1 and (1.7) if we choose $\varphi = B\xi$ and $\varphi^N = \iota^N B^N \rho^N \xi$. □

In case of the spline scheme discussed in § 3 we have uniform convergence of the transfer functions (3.2) on compact intervals to the transfer function of the delay system (1.1) [5].

REFERENCES

[1] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear System Theory*, Lecture Notes in Control and Information Sciences Vol. 8, Springer-Verlag, New York, 1978.

[2] J. H. CURTISS, *Introduction to Functions of a Complex Variable*, Marcel Dekker, New York, 1978.

[3] J. S. GIBSON, *Linear quadratic optimal control of hereditary systems: infinite-dimensional Riccati equations and numerical approximation*, SIAM J. Control Optim., 21 (1983), pp. 95-139.

[4] F. KAPPEL AND D. SALAMON, *Spline approximation for retarded systems and the Riccati equation*, SIAM J. Control Optim., 25 (1987), pp. 1082-1117.

[5] ———, *On the stability properties of spline approximations for retarded systems*, SIAM J. Control Optim., 27 (1989), pp. 407-431.

[6] J. ZABCZYK, *Remarks on the algebraic Riccati equation*, Appl. Math. Optim., 2 (1976), pp. 251-258.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 1148-1161, September 1990

007

THE EXPONENTIAL FORMULA FOR THE REACHABLE SET OF A LIPSCHITZ DIFFERENTIAL INCLUSION*

PETER R. WOLENSKI†

Abstract. The main goal of this paper is to prove a formula for the reachable set of a Lipschitz differential inclusion with convex values. The formula involves a Kuratowski limit of sets that resembles a standard approach of defining the exponential of a matrix—this explains the title. The proof of the main theorem partially relies on a C^1 approximation result due to Filippov, for which a new proof is given. A new approach of characterizing the value function associated with a Mayer optimal control problem is given as an application.

Key words. differential inclusions, exponential formula, reachable sets, time discretization of differential inclusions

AMS(MOS) subject classifications. 34A60, 49E10, 34A45

1. Introduction. A differential inclusion generalizes an ordinary differential equation by permitting set-valued right-hand sides. That is, an (autonomous) differential inclusion has the form

$$\begin{aligned} &x(\cdot) \text{ absolutely continuous on } [0, T] \text{ into } X \\ &\dot{x}(t) \in F(x(t)) \quad \text{a.e. } t \in [0, T] \\ &x(0) = \xi, \end{aligned} \tag{1.1}$$

where $T > 0$, $X \subseteq \mathfrak{R}^n$ is open, $\xi \in X$, $F: \mathfrak{R}^n \rightrightarrows \mathfrak{R}^n$ is a multifunction (or set-valued map), and $\dot{x}(t)$ denotes the derivative of $x(\cdot)$ with respect to t. As is well known [cf. 2, § 1.3], standard control models with feedback can be transformed into the framework of (1.1). The main advantage of such a transformation is its mathematical simplicity, since the control parameters do not explicitly appear. This advantage will be evident in the present paper, in which we prove a differential inclusion analogue of Euler's classical technique for solving ordinary differential equations by successive approximation. The results are then applied to characterize the value function associated with a Mayer optimal control problem.

Differential inclusions receive a broad treatment in Aubin and Cellina [1], to which we also refer the reader for early references in this subject. See Clarke [2, § 3.1] for a concise exposition of the basic properties of multifunctions and differential inclusions.

The main focus of our analysis of (1.1) involves the reachable set $R^{(T)}(\xi)$, which is defined by

$$R^{(T)}(\xi) := \{x(T): x(\cdot) \text{ solves (1.1)}\}. \tag{1.2}$$

Reachable sets seem to have better properties than the set of trajectories to (1.1). This is illustrated in [11] where a uniqueness theorem for differential inclusions, analogous to the classical ordinary differential equation (o.d.e.) uniqueness theorem, is proven in terms of the semigroup properties in t of the multifunction $\xi \rightrightarrows R^{(T)}(\xi)$. We point out that in o.d.e. theory, the distinction between the solutions (which are functions of t) and the end points of these solutions (which are points of $\mathfrak{R}^n$) is generally not

* Received by the editors November 4, 1988; accepted for publication (in revised form) June 21, 1989.
† Centre de Recherches Mathématiques, Université de Montreal, Montreal, Quebec H3C 3J7, Canada.

necessary. This gives a flavor to differential inclusion theory that has no meaningful counterpart in o.d.e. theory.

The main result of this paper is that the formula

$$R^{(T)}(\xi) = \lim_{N\to\infty} \left(I + \frac{T}{N} F\right)^N (\xi) \tag{1.3}$$

holds under local Lipschitz and convexity assumptions on F. Here the power of $(I+(T/N)F)$ is that of composition of multifunctions, and the limit is a set limit in the sense of Kuratowski. For obvious reasons, we call (1.3) the exponential formula.

An important feature of the formula (1.3) is that solutions to (1.1) need not be invoked to determine the points in $R^{(T)}(\xi)$. This was also the case in a result by Vinter [9], where a certain condition is shown to determine whether a set intersects $R^{(T)}(\xi)$. The methods employed are quite abstract and involve an auxiliary optimization problem that resembles a Hamilton–Jacobi type inequality and tools of convex analysis.

The idea of applying a Kuratowski limit to the sequence $(I+(T/N)F)^N(\xi)$ appears in Rockafellar [6], but only as a heuristic tool in generating one-parameter semigroups of convex processes. Whether the heuristic argument could be formalized was not resolved, and is still not for an arbitrary convex process (which is apparently important in economic modeling). However we prove (1.3) under quite general assumptions on F.

Discretization of differential inclusions has also been explored by Taubert [8]. The assumption on F is merely upper semicontinuity, hence it is weaker than our Lipschitz assumption. The main result is that at least one sequence of "discrete" trajectories to the associated discrete differential inclusion converges to a solution of (1.1); and as a partial converse, if there is a *unique* solution to (1.1), then every such discretization converges to it. Under the stronger Lipschitz assumption, our result (1.3) implies that the set of all endpoints of discrete trajectories actually converge to the entire reachable set.

A paper by Dontchev and Farkhi [4] has recently been brought to our attention[1] which contains some results very similar to ours. The focus of these results is to approximate the trajectories rather than the reachable set, and is thus of a somewhat different flavor.

The plan for the rest of the paper is as follows: preliminaries are in § 2; § 3 is devoted to a straightforward proof of a C^1 approximation result due to Filippov (this also has considerable independent interest); § 4 contains the precise statement and proof of the exponential formula; § 5 has two related results; the time-dependent case is stated in § 6; § 7 consists of simple proofs based on the exponential formula of two well known theorems; and finally, § 8 contains an application to the Mayer optimal control problem. Further comments on this application will be given there.

2. Preliminaries. Throughout the rest of the paper, T will be a nonnegative real number. The interval $[0, T]$ should be thought of as a time interval.

Most of the notation is standard. The absolutely continuous functions on $[0, T]$ are denoted by $AC[0, T]$, and the continuously differentiable functions by $C^1[0, T]$. If $x(\cdot)$ is continuous on $[0, T]$, then $\|x\|$ will denote its sup norm.

For $a \in \mathfrak{R}^n$ and $A \subseteq \mathfrak{R}^n$, the distance from a to A is defined by $\operatorname{dist}(a, A) = \inf\{|a-a'|: a' \in A\}$. If A_0 and A_1 are two nonempty compact subsets of $\mathfrak{R}^n$, the Hausdorff distance is denoted by $\operatorname{dist}_H(A_0, A_1)$ and equals the smallest δ for which

[1] The author wishes to thank V. M. Veliov for this and other helpful conversations.

$\sup_{a_1 \in A_1} \operatorname{dist}(a_1, A_0) \leqq \delta$ and $\sup_{a_0 \in A_0} \operatorname{dist}(a_0, A_1) \leqq \delta$ both hold. It is easy to check that $\operatorname{dist}_H(\cdot, \cdot)$ is a metric on the nonempty compact subsets of $\Re^n$.

In the following, we will frequently use Kuratowski limits of sets. If $\{A_j\}_{j=1}^{\infty}$ is a sequence of subsets of $\Re^n$, define the lim sup and lim inf of $\{A_j\}_{j=1}^{\infty}$ by

$$\limsup_{j\to\infty} A_j = \left\{ a: \liminf_{j\to\infty} \operatorname{dist}(a, A_j) = 0 \right\} \tag{2.1}$$

$$\liminf_{j\to\infty} A_j = \left\{ a: \limsup_{j\to\infty} \operatorname{dist}(a, A_j) = 0 \right\}. \tag{2.2}$$

If lim sup A_j equals lim inf A_j, we say that the limit exists and write $\lim_{j\to\infty} A_j$ for the common value. Note that lim sup A_j and lim inf A_j are always closed sets. If each of A_j and A are nonempty and compact and are contained in a given bounded set, then it is immediate from (2.1) and (2.2) that $A = \lim_{j\to\infty} A_j$ if and only if $\operatorname{dist}_H(A_j, A) \to 0$ as $j \to \infty$.

Let $G: \Re^m \rightrightarrows \Re^n$ be a given multifunction. Then G is *upper semicontinuous* at $\xi_0 \in \Re^n$ if $G(\xi_0) \supseteq \limsup_{j\to\infty} G(\xi_j)$ for all sequences $\{\xi_j\}$ satisfying $\xi_j \to \xi_0$. G is *lower semicontinuous* at ξ_0 if $G(\xi_0) \subseteq \liminf_{j\to\infty} G(\xi_j)$ for all $\{\xi_j\}$ with $\xi_j \to \xi_0$. G is continuous if it is both upper and lower semicontinuous. If G also has compact values on a subset X of $\Re^n$, then G is Lipschitz of order $\lambda > 0$ on X if $\operatorname{dist}_H(G(\xi), G(\xi')) \leqq \lambda |\xi - \xi'|$ for all $\xi, \xi' \in X$. We say G is locally Lipschitz on X if G is Lipschitz on each compact subset of X.

Again consider the differential inclusion (1.1). We define the *solution set*, or set of trajectories for F, by

$$S^{(T)}(\xi) := \{x(\cdot): x(\cdot) \text{ satisfies (1.1) in } X\}. \tag{2.3}$$

Let $X \subseteq \Re^n$ be open and $\xi \in X$. The *escape time* $T_X(\xi)$ *from* X is defined as the smallest T for which there exists $x(\cdot) \in S^{(T)}(\xi)$ so that as $t \uparrow T$, either $|x(t)| \to +\infty$ or $x(t)$ approaches the boundary of X. It can be shown that if F has compact values and is locally Lipschitz on X, then

$$T_X(\xi) = \sup \left\{ T: \operatorname{cl} \bigcup_{0 \leqq t \leqq T} R^{(t)}(\xi) \text{ is compact} \right\}. \tag{2.4}$$

If F also has convex values the "cl" can be removed from the set in (2.4). A proof of this, as well as other information regarding escape times, can be found in [11].

One of the most useful tools in the theory of differential inclusions is the following theorem due to Filippov. Our statement assumes more than is necessary in that F has compact values, but this will suffice for our purposes. If $x(\cdot) \in AC[0, T]$, define

$$\rho(x) := \int_0^T \operatorname{dist}(\dot{x}(t), F(x(t)))\, dt.$$

THEOREM 2.1 (Filippov [5]). *Let $F: \Re^n \rightrightarrows \Re^n$ be a compact-valued multifunction and $x(\cdot)$ an absolutely continuous function on $[0, T]$. Suppose there exists $\delta > 0$ and a set K for which F is Lipschitz of order $\lambda > 0$ on K and $\{\xi: |\xi - x(t)| \leqq \delta$ for some $t \in [0, T]\}$ is contained in K. If $\rho(x) < \delta e^{-\lambda T}$, then there exists $\bar{x} \in S^{(T)}(x(0))$ satisfying $\|x - \bar{x}\| < \rho(x) e^{\lambda T}$.*

A proof of Theorem 2.1 is, for example, in Carke [2, p. 115].

We close this section with a simple technical lemma.

LEMMA 2.2. *Suppose* α, β, $m_1, m_2, \cdots, m_N$ *are real constants satisfying* $m_{j+1} = \alpha + \beta m_j$ *for* $j = 1, \cdots, N$, *then*

$$m_N = \begin{cases} \alpha\left(\dfrac{1-\beta^N}{1-\beta}\right) + \beta^N m_0 & \text{if } \beta \neq 1 \\ N\alpha + m_0 & \text{if } \beta = 1. \end{cases}$$

The proof is elementary.

3. Approximation by C^1 trajectories. This section is devoted to the proof of a special case of Filippov [3, Thm. 6]. We first motivate the result. Suppose the values of F in (1.1) consist of singleton sets, say $F(x) = \{f(x)\}$, where $f: \Re^n \to \Re^n$ is a Lipschitz function. Then $S^{(T)}(\xi) = \{x(\cdot)\}$, where $x(\cdot)$ satisfies $\dot{x}(t) = f(x,(t))$ for all $t \in [0, T]$. Note that $x(\cdot)$ is not merely absolutely continuous, but an element of $C^1[0, T]$ as well. Is there an analogue of the additional regularity of solutions when F is multivalued? Of course, in the more general context $S^{(T)}(\xi)$ will not consist entirely of C^1 elements, but it is reasonable to ask if $C^1[0, T] \cap S^{(T)}(\xi)$ is dense in $S^{(T)}(\xi)$, say in the sup norm topology. We show that this is indeed the case when F is assumed to have convex values and to be locally Lipschitz. Filippov's original result is more general in that unbounded values of F are permitted and convexity is replaced by "uniformly locally connected." Under the simplified assumptions we provide a simpler proof.

THEOREM 3.1 (Filippov [5]). *Suppose* $X \subseteq \Re^n$ *is open and* $F: X \rightrightarrows \Re^n$ *is a multifunction with nonempty, convex, and compact values. Assume* F *is locally Lipschitz on* X. *Let* $T > 0$ *and suppose* $x(\cdot) \in S^{(T)}(x(0))$. *Then for each* $\varepsilon > 0$ *there exists* $\bar{x}(\cdot) \in S^{(T)}(x(0)) \cap C^1[0, T]$ *satisfying* $\|x - \bar{x}\| < \varepsilon$.

Much of the work in our proof of Theorem 3.1 is contained in the proof of the following proposition.

PROPOSITION 3.2. *Let* X *and* F *be as in the statement of Theorem* 3.1. *Let* $T > 0$. *Suppose* $y(\cdot) \in C^1[0, T]$, $\delta > 0$, *and* $K \subseteq \Re^n$ *is compact so that* $\{\xi: |\xi - y(t)| \leqq \delta \text{ for some } 0 \leqq t \leqq T\} \subseteq K \subseteq X$. *Let* $\lambda > 0$ *be a Lipschitz constant for* F *on* K. *Assume further that* $\rho(y) < \delta e^{-\lambda T}$. *Then there exists* $\bar{y} \in S^{(T)}(y(0)) \cap C^1[0, T]$ *satisfying* $\|y - \bar{y}\| < \rho(y)\, e^{\lambda T}$.

One may note that Proposition 3.2 resembles the oft-quoted Theorem 2.1 of Filippov, and those familiar with the proof of Theorem 2.1 will see that the proof of Proposition 3.2 is basically the same. The difference here is that by assuming the given function $y(\cdot)$ is C^1, one may conclude that the trajectory $\bar{y}$ is also C^1.

The proof of Proposition 3.2 requires the following lemma. If $A \subseteq \Re^n$ is a closed convex set and $a \in \Re^n$, we denote by proj (a, A) the unique element in A closest to a.

LEMMA 3.3. *Suppose* $G: [0, T] \rightrightarrows \Re^n$ *is a continuous multifunction with nonempty, closed, and convex values on* $[0, T]$. *Also suppose* $v: [0, T] \to \Re^n$ *is a continuous function. Then the function* $t \to \text{proj}\,(v(t), G(t))$ *is continuous on* $[0, T]$.

Proof. Set $p(t) = \text{proj}\,(v(t), G(t))$. Let $t_0 \in [0, T]$ and $\{t_j\}_{j \geqq 1} \subseteq [0, T]$ with $t_j \to t_0$ as $j \to \infty$. Note that since G is continuous, the sequence $\{p(t_j)\}_{j \geqq 1}$ is bounded. Let $\bar{p}$ be any cluster point of $\{p(t_j)\}_{j \geqq 1}$. Passing to a subsequence if necessary, but without renaming it, we assume $p(t_j) \to \bar{p}$ as $j \to \infty$. It suffices to show $p(t_0) = \bar{p}$.

Since G is upper semicontinuous at t_0, we have $\bar{p} \in G(t_0)$. Since G is lower semicontinuous at t_0, there exists $q_j \in G(t_j)$ so that $q_j \to p(t_0)$. Hence we have

$$\begin{aligned} |v(t_0) - \bar{p}| &= \lim_{j \to \infty} |v(t_j) - p(t_j)| \\ &\leqq \lim_{j \to \infty} |v(t_j) - q_j| \qquad (3.1) \\ &= |v(t_0) - p(t_0)|. \end{aligned}$$

But now $p(t_0)$ is the unique element in $G(t_0)$ closest to $v(t_0)$, so (3.1) implies that $p(t_0) = \bar{p}$. □

Proof of Proposition 3.2. The C^1 function y is given satisfying $\rho(y) < \delta e^{-\lambda T}$. By the lemma, $v_0(t) := \text{proj}\,(\dot{y}(t), F(y(t)))$ is continuous on $[0, T]$. Set

$$y_1(t) = y(0) + \int_0^t v_0(s)\, ds.$$

Then $y_1(\cdot) \in C^1[0, T]$ with $\dot{y}_1(t) \in F(y(t))$ for all t. Also, $y_1(t)$ is contained in K for all t because

$$\begin{aligned} |y_1(t) - y(t)| &\leqq \int_0^t |v_0(s) - \dot{y}(s)|\, ds \\ &= \int_0^t \text{dist}\,(\dot{y}(s), F(y(s)))\, ds \qquad (3.2) \\ &\leqq \rho(y) < \delta. \end{aligned}$$

Set $y_0(\cdot) = y(\cdot)$. Inductively, suppose $k \geqq 1$ and C^1 functions $\{y_j\}_{j=0}^k$ have been chosen so that (3.3)–(3.6) hold for $0 \leqq t \leqq T$ and $1 \leqq j \leqq k$:

$$\dot{y}_j(t) \in F(y_{j-1}(t)), \tag{3.3}$$

$$|\dot{y}_j(t) - \dot{y}_{j-1}(t)| \leqq \rho(y) \frac{\lambda^{j-1} t^{j-2}}{(j-2)!}, \tag{3.4}$$

$$|y_j(t) - y_{j-1}(t)| \leqq \rho(y) \frac{(\lambda t)^{j-1}}{(j-1)!}, \tag{3.5}$$

$$y_j(t) \in K. \tag{3.6}$$

When $j = 1$, (3.3) is obvious, (3.4) is vacuous, and (3.5) and (3.6) follow immediately from (3.2).

We proceed by defining $v_{k+1}(t) = \text{proj}\,(\dot{y}_k(t), F(y_k(t))$ and $y_{k+1}(t) = y(0) + \int_0^t v_{k+1}(s)\, ds$. By Lemma 3.3, v_{k+1} is continuous and hence $y_{k+1}(\cdot) \in C^1[0, T]$. Let $t \in [0, T]$. It is immediate that $\dot{y}_{k+1}(t) \in F(y_k(t))$. We have

$$\begin{aligned} |\dot{y}_{k+1}(t) - \dot{y}_k(t)| &= \text{dist}\,(\dot{y}_k(t), F(y_k(t)) \\ &\leqq \lambda |y_{k-1}(t) - y_k(t)| \quad \text{(by (3.3) and Lipschitz property)} \qquad (3.7) \\ &\leqq \rho(y) \frac{\lambda^k t^{k-1}}{(k-1)!} \qquad \text{(by (3.5))}. \end{aligned}$$

This shows that (3.4) holds for $j = k+1$. We also have

$$\begin{aligned} |y_{k+1}(t) - y_k(t)| &\leqq \int_0^t |\dot{y}_{k+1}(s) - \dot{y}_k(s)|\, ds \\ &\leqq \rho(y) \int_0^t \frac{\lambda^k s^{k-1}}{(k-1)!}\, ds \quad \text{(by (3.7))} \qquad (3.8) \\ &\leqq \rho(y) \frac{(\lambda t)^k}{k!}, \end{aligned}$$

which shows that (3.5) holds for $j=k+1$. Finally $y_{k+1}(t)\in K$ because

$$
\begin{aligned}
|y_{k+1}(t)-y(t)| &\leqq \sum_{j=0}^{k} |y_{j+1}(t)-y_j(t)| \\
&\leqq \rho(y)\sum_{j=0}^{k}\frac{(\lambda t)^j}{j!} \quad \text{(by (3.5) and (3.8))} \\
&\leqq \rho(y)\, e^{\lambda T} \leqq \delta.
\end{aligned} \tag{3.9}
$$

The induction is now complete.

We have constructed a subsequence $\{y_j\}_{j=1}^{\infty}$ of C^1 functions lying in K. By (3.4) and (3.5) this sequence is Cauchy in the Sobolev norm on $C^1[0, T]$, and hence there exists $\bar{y}\in C^1[0, T]$ for which $\dot{y}_j\to\dot{\bar{y}}$ and $y_j\to\bar{y}$, both uniformly on $[0, T]$. Moreover, for each $0\leqq t\leqq T$,

$$
\dot{\bar{y}}(t)=\lim_{j\to\infty}\dot{y}_j(t)\in\lim_{j\to\infty}F(y_{j-1}(t))=F(\bar{y}(t)),
$$

and so $\bar{y}(\cdot)\in S^{(T)}(y(0))$. From (3.9) we obtain the desired estimate of the difference between $\bar{y}$ and y:

$$
\|\bar{y}-y\|=\sup_{0\leqq t\leqq T}\lim_{j\to\infty}|y_j(t)-y(t)|\leqq\rho(y)\, e^{\lambda T}. \qquad \square
$$

Proof of Theorem 3.1. We are now given $x(\cdot)\in S^{(T)}(x(0))$ and $\varepsilon>0$. We must show there exists $\bar{x}(\cdot)\in S^{(T)}(x(0))\cap C^1[0, T]$ with $\|x-\bar{x}\|<\varepsilon$. Without loss of generality, we may assume ε is sufficiently small so that there exists a compact set K satisfying $\{\xi: |\xi-x(t)|\leqq\varepsilon \text{ for some } 0\leqq t\leqq T\}\subseteq K\subseteq X$. Let $\lambda\geqq 1$ be a Lipschitz constant for F on K and let $r=\sup\{|v|: v\in F(K)\}<\infty$. Note that $|\dot{x}(t)|\leqq r$ for almost all $0\leqq t\leqq T$.

By Lusin's Theorem (cf. [7, p. 46]), there exists a continuous function $z(\cdot)$ on $[0, T]$ and a Borel set $J\subseteq[0, T]$ so that $z(t)=\dot{x}(t)$ for $t\notin J$, $\|z\|\leqq r$, and $m(J)<\varepsilon/4\lambda r(1+T)\, e^{\lambda T}$ (where m is Lebesque measure on $[0, T]$).

Define $y(t)=x(0)+\int_0^t z(s)\, ds$. Then $y(\cdot)\in C^1[0, T]$ and the following holds for all $0\leqq t\leqq T$:

$$
\begin{aligned}
|y(t)-x(t)| &\leqq \int_J |z(s)-x(s)|\, ds \\
&\leqq 2rm(J) \\
&\leqq \frac{\varepsilon}{2\lambda\, e^{\lambda T}(1+T)} \leqq \frac{\varepsilon}{2}.
\end{aligned} \tag{3.10}
$$

In particular (3.10) implies that $\{\xi: |\xi-y(t)|\leqq(\varepsilon/2)\}$ is contained in K. Next we estimate $\rho(y)$:

$$
\begin{aligned}
\rho(y) &= \int_0^T \operatorname{dist}(\dot{y}(t), F(y(t)))\, dt \\
&\leqq \int_{[0,T]\setminus J} \operatorname{dist}_H(F(x(t)), F(y(t)))\, dt + \int_J \operatorname{dist}(\dot{y}(t), F(y(t)))\, dt \\
&\leqq \lambda\int_0^T |x(t)-y(t)|\, dt + 2rm(J) \\
&< \frac{\varepsilon}{2}e^{-\lambda T} \quad \text{(by (3.10))}.
\end{aligned}
$$

As a consequence of Proposition 3.2 applied to $y(\cdot)$, $\delta = \varepsilon/2$, and K, there exists $\bar{x}(\cdot) \in S^{(T)}(x(0)) \cap C^1[0, T]$ so that $\|y - \bar{x}\| < \varepsilon/2$. Finally, we conclude that $\|x - \bar{x}\| \leqq \|x - y\| + \|y - \bar{x}\| < \varepsilon$. □

4. The exponential formula. We now come to the main result. If G_0 and G_1 are two multifunctions from $\mathfrak{R}^n$ into $\mathfrak{R}^n$, we define the composition $G_0 \circ G_1 : \mathfrak{R}^n \rightrightarrows \mathfrak{R}^n$ of G_1 with G_0 by $G_0 \circ G_1(x) = \{z$: there exists $y \in G_1(x)$ with $z \in G_0(y)\}$. If G_0 is composed with itself N times, we write G_0^N for the resulting multifunction.

THEOREM 4.1. *Suppose $X \subseteq \mathfrak{R}^n$ is open and $F : \mathfrak{R}^n \rightrightarrows \mathfrak{R}^n$ is a multifunction. Assume F has nonempty compact values on X and is locally Lipschitz on X. Fix $\xi \in X$.*

(i) *For $0 \leqq T < T_X(\xi)$, one has*

$$\limsup_{N \to \infty} \left(I + \frac{T}{N} F \right)^N (\xi) \subseteq \operatorname{cl} R^{(T)}(\xi). \tag{4.1}$$

(ii) *If, in addition, F is assumed to have convex values, then for all $T \geqq 0$, one has*

$$R^{(T)}(\xi) \subseteq \liminf_{N \to \infty} \left(I + \frac{T}{N} F \right)^N (\xi). \tag{4.2}$$

Proof. (i) Suppose $0 \leqq T < T_X(\xi)$. Let $K = \operatorname{cl} \bigcup_{0 \leqq t \leqq T} R^{(t)}(\xi)$. By definition of $T_X(\xi)$, K is compact, so there exists $\delta > 0$ so that $K + \delta B \subseteq X$, where B is the closed unit ball. Set $r = \sup\{|v|: v \in F(K + \delta B)\}$ and choose $\lambda > 0$ to be a Lipschitz constant for F on $K + \delta B$.

Now let $\varepsilon > 0$. We show that for all large N satisfying

$$\frac{T}{N} \leqq \min\left\{ \frac{\varepsilon}{\lambda r T e^{\lambda T}}, \frac{\delta}{2r} \right\},$$

the inclusion

$$\left(I + \frac{T}{N} F \right)^j (\xi) \subseteq R^{(jT/N)}(\xi) + \varepsilon B \tag{4.3}$$

holds for all $j = 0, 1, \cdots, N$. Since ε is arbitrarily small, by setting $j = N$ in (4.3), we can then immediately conclude that (4.1) holds.

Let N be as above. To simplify the notation, set $h = T/N$ and $t_j = jh$ for $j = 0, 1, \cdots, N$. We prove (4.3) by induction on j. The case $j = 0$ is trivial. For the induction hypothesis, suppose (4.3) holds for all i, $0 \leqq i \leqq j < N$. Let $y_{j+1} \in (I + hF)^{j+1}(\xi)$. There exists $y_0 = \xi, y_1, \cdots, y_j$ and $u_0, \cdots, u_j$ so that for $0 \leqq i \leqq j$, we have

$$u_i \in F(y_i) \quad \text{and} \quad y_{i+1} = y_i + h u_i.$$

Note that when $0 \leqq i \leqq j$, (4.3) implies $y_i \in K + (\delta/2)B$ and thus $|u_i| \leqq r$. Let $x(\cdot)$ be defined on $[0, t_{j+1}]$ as the piecewise linear interpolation of $\{y_i\}_{i=0}^{j+1}$ equally spaced on $[0, t_{j+1}]$. That is,

$$x(t) = y_i + (t - t_i) u_i \quad \text{if } t_i \leqq t \leqq t_{i+1}.$$

The range of $x(\cdot)$ lies within $K+\delta B$ because $y_i+(t-t_i)u_i \in K+(\delta/2)B+hrB \subseteq K+\delta B$. Hence we have

$$
\begin{aligned}
\rho(x) &= \int_0^{t_{j+1}} \operatorname{dist}(\dot{x}(t), F(x(t)))\, dt \\
&\leqq \sum_{i=0}^{j} \int_{t_i}^{t_{i+1}} \operatorname{dist}_H(F(y_i), F(x(t)))\, dt \qquad (4.4) \\
&\leqq \lambda \sum_{i=0}^{j} \int_{t_i}^{t_{i+1}} |y_i - x(t)|\, dt \quad \text{(by Lipschitz property of } F \text{ on } K+\delta B) \\
&\leqq \lambda T r h.
\end{aligned}
$$

By the Filippov result Theorem 2.1, we may conclude from (4.4) that

$$
\begin{aligned}
\operatorname{dist}_H(y_{j+1}, R^{(t_{j+1})}(\xi)) &\leqq \rho(x)\, e^{\lambda T} \\
&< \lambda T r\, e^{\lambda T} h \\
&< \varepsilon.
\end{aligned}
$$

Thus (4.3) holds for all $j=0, 1, \cdots, N$, and the proof of (i) is complete.

(ii) The values of F are now assumed to be convex, and thus Theorem 3.1 is applicable. Let $x(\cdot) \in C^1[0, T] \cap S^{(T)}(\xi)$. A consequence of Theorem 3.1 is that to prove (4.2), it suffices only to show $x(T) \in \liminf_{j\to\infty} (I+(T/N)F)^N(\xi)$.

Denote the range of $x(\cdot)$ by K and choose $\delta>0$ so that $K+\delta B \subseteq X$. Let $\lambda>0$ be a Lipschitz constant for F on $K+\delta B$. For each integer N, define

$$
\varepsilon_N = \sup_{j=0,1,\cdots,N} \left| \frac{x(t_{j+1})-x(t_j)}{h} - \dot{x}(t_j) \right|, \tag{4.5}
$$

where $h=(T/N)$ and $t_j=jh$ for $j=0, 1, \cdots, N$. It is immediate that $\varepsilon_N \to 0$ as $N\to\infty$ because $x(\cdot) \in C^1[0, T]$. If N is sufficiently large so that $\varepsilon_N < (\delta\lambda/e^{\lambda T}-1)$, we claim that

$$
x(T) \in \left(I+\frac{T}{N}F\right)^N(\xi) + \frac{\varepsilon_N}{\lambda}(e^{\lambda T}-1)B. \tag{4.6}
$$

To prove the claim (4.6), we start by letting $y_0=\xi$, $u_0=\dot{x}(t_0)$, and $m_0=0 \in \Re$. Having chosen y_j, u_j, and m_j, let $y_{j+1}=y_j+hu_j$, $u_{j+1}=\operatorname{proj}(\dot{x}(t_{j+1}), F(y_{j+1}))$ and $m_{j+1}=(1+\lambda h)m_j+1$. Note that $m_j \leqq m_{j+1}$ for each j, and by Lemma 2.2 (with $\alpha=1$, $\beta=(1+\lambda h)$), we have

$$
m_N = \frac{1}{\lambda h}((1+\lambda h)^N - 1) \leqq \frac{1}{\lambda h}(e^{\lambda T}-1). \tag{4.7}
$$

Inductively, suppose for $0 \leqq j < N$, the estimate

$$
|y_j - x(t_j)| \leqq h\varepsilon_N m_j \tag{4.8}
$$

holds. When $j=0$, (4.8) is trivial. We have from (4.7) that $h\varepsilon_N m_j \leqq \varepsilon_N(e^{\lambda T}-1)/\lambda$, which by the choice of N is $\leqq\delta$. Hence (4.8) implies $y_j \in K+\delta B$. By the Lipschitz property of F on $K+\delta B$ and the choice of u_j, we have

$$
\begin{aligned}
|u_j - \dot{x}(t_j)| &\leqq \lambda |y_j - x(t_j)| \\
&\leqq \lambda h \varepsilon_N m_j \quad \text{(by (4.8))}.
\end{aligned} \tag{4.9}
$$

Therefore

$$\begin{aligned}|y_{j+1}-x(t_{j+1})| &\leqq |y_j-x(t_j)|+h|u_j-\dot{x}(t_j)|+|x(t_j)+h\dot{x}(t_j)-x(t_{j+1})| \\ &\leqq h\varepsilon_N m_j+\lambda h^2\varepsilon_N m_j+h\varepsilon_N \quad \text{(by (4.8), (4.9), and (4.5))} \\ &= h\varepsilon_N m_{j+1}.\end{aligned}$$

Consequently the estimate (4.8) holds for $j+1$. When $j=N$, (4.8) and (4.7) combine to give us

$$|y_N-x(T)|\leqq \frac{\varepsilon_N}{\lambda}(e^{\lambda T}-1). \tag{4.10}$$

Finally the claim (4.6) follows from (4.10) and the observation that $y_N \in (I+(T/N)F)^N(\xi)$.

Now that (4.6) is verified for all large N, the conclusion (4.2) follows from (4.6) be letting $N\to\infty$. □

We record the exponential formula in the next corollary. This is an immediate consequence of Theorem 4.1.

COROLLARY 4.2. *Suppose $X\subseteq \mathfrak{R}^n$ is open and $F:\mathfrak{R}^n \rightrightarrows \mathfrak{R}^n$ is a multifunction. Assume F has nonempty, compact, convex values on X and is locally Lipschitz on X. Then for all $\xi\in X$ and $0\leqq T<T_X(\xi)$, we have*

$$R^{(T)}(\xi)=\lim_{N\to\infty}\left(I+\frac{T}{N}F\right)^N(\xi).$$

5. Related results. In this section, we state some results whose proofs require only minor modifications of the proof of Theorem 4.1.

COROLLARY 5.1. *Suppose X and F are as in Corollary* 4.2, *and let $\xi\in X$ and $0\leqq t\leqq T<T_X(\xi)$. If $\{N_l\}_{l=1}^{\infty}$ and $\{j_l\}_{l=1}^{\infty}$ are a pair of sequences of nonnegative integers with $N_l\to\infty$, $0\leqq j_l\leqq N_l$, and $(j_lT/N_l)\to t$, then*

$$R^{(t)}(\xi)=\lim_{l\to\infty}\left(I+\frac{T}{N_l}F\right)^{j_l}(\xi). \tag{5.1}$$

Proof. The multifunction $s\rightrightarrows R^{(s)}(\xi)$ is continuous on $[0, T]$ (for example, see [11, Lemma 5.2]). Hence from (4.3), one deduces that

$$\limsup_{l\to\infty}\left(I+\frac{T}{N_l}F\right)^{j_l}(\xi)\subseteq R^{(t)}(\xi). \tag{5.2}$$

A consequence of (4.8) is (with l sufficiently large) that

$$\begin{aligned}R^{(j_lT/N_l)}(\xi) &\subseteq \left(I+\frac{T}{N_l}F\right)^{j_l}(\xi)+h\varepsilon_{N_l}m_j \\ &\subseteq \left(I+\frac{T}{N_l}F\right)^{j_l}(\xi)+\varepsilon_{N_l}\frac{1}{\lambda}(e^{\lambda T}-1).\end{aligned} \tag{5.3}$$

Letting $l\to\infty$ in (5.3) implies

$$R^{(t)}(\xi)\subseteq \liminf_{l\to\infty}\left(I+\frac{T}{N_l}F\right)^{j_l}(\xi). \tag{5.4}$$

Combining (5.4) and (5.2) begets (5.1). □

The exponential formula (Corollary 4.2) is concise, but in applications it may be desirable to have a more flexible approximation to the reachable set. For example, in a numerical approximation to an optimization problem of the type (1.2), it may not be practical to take a uniform discretization of the time interval. The next corollary

indicates that one may partition $[0, T]$ in an arbitrary manner provided the width of the largest subinterval goes to zero. In other words, one may say that discrete approximation of Lipschitz differential inclusions is robust.

To state the result, we need a few more definitions. If $P = \{t_0, \cdots, t_N\}$ is a partition of $[0, T]$ (that is, $0 = t_0 < t_1 < \cdots < t_N = T$), define $|P| = \sup_{0 \leqq j \leqq N-1} |t_{j+1} - t_j|$. If $\{H_j\}_{j=1}^N$ is a collection of multifunctions from $\mathfrak{R}^n$ to $\mathfrak{R}^n$, define the multifunction product by $(\prod_{j=1}^N H_j)(\xi) = (H_N \circ H_{N-1} \circ \cdots \circ H_1)(\xi)$.

COROLLARY 5.2. *Suppose X and F are as in Corollary* 4.2, *and let $\xi \in X$ and $0 \leqq T < T_X(\xi)$. Then for any sequence of partitions $P_k = \{t_0^k, t_1^k, \cdots, t_{N_k}^k\}$ of $[0, T]$ with $|P_k| \to 0$ as $k \to \infty$, we have*

$$R^{(T)}(\xi) = \lim_{k \to \infty} \left(\prod_{j=0}^{N_k - 1} (I + (t_{j+1}^k - t_j^k)F) \right)(\xi).$$

The proof is left to the reader. One needs only to mimic the steps in the proof of Theorem 4.1 by replacing h by $h_j^k := t_{j+1}^k - t_j^k$, etc.

The last result of this section will be used in §§ 7 and 8. It says that the limit in Corollary 5.2 is "uniform" over ξ in a compact set. For $K \subseteq X$ compact, define $T_X(K) = \inf\{T_X(\xi): \xi \in K\}$. It is shown in [11] that $T_X(K) > 0$. Again it is convenient to return to partitions of $[0, T]$ with elements of equal length, but the result easily extends to arbitrary partitions.

PROPOSITION 5.3. *Suppose X and F are as in Corollary* 4.2. *Let $\varepsilon > 0$, $K_0 \subseteq X$ compact, and $0 < T < T_X(K_0)$. Then there exists $N_0 > 0$ independent of $\xi \in K_0$ so that for each $N \geqq N_0$, $j = 0, 1, \cdots, N$, and $\xi \in K_0$, we have*

$$\operatorname{dist}_H (R^{(t_j)}(\xi), (I + hF)^j(\xi)) < \varepsilon, \tag{5.5}$$

where $h = T/N$ and $t_j = jh$.

Proof. One needs only to check that N in the proof of Theorem 4.1 can be chosen independent of $\xi \in K_0$. To this end, define

$$K = \operatorname{cl} \cup \{R^{(t)}(\xi): 0 \leqq t \leqq T,\ \xi \in K_0\}.$$

Then K is compact. Choose the constants δ, r, and λ as in the proof of Theorem 4.1 using this K. Then for all large N and $0 \leqq j \leqq N$, one has that (4.3) holds for each $\xi \in K_0$.

It remains only to show that ε_N defined in (4.6) satisfies $\varepsilon_N \to 0$ as $N \to \infty$ independent of $\xi \in K_0$. Since $x(t_j) \in K$ and $\dot{x}(t_j) \in F(x(t_j))$, the estimate

$$\varepsilon_N \leqq \sup_{\eta \in K} \operatorname{dist}_H \left(\frac{R^{(h)}(\eta) - \eta}{h}, F(\eta) \right) \tag{5.6}$$

holds. By [11, Thm. 3.1(d)], the right side of (5.6) approaches 0 as $N \to \infty$. Therefore the term $(\varepsilon_N/\lambda)(e^{\lambda T} - 1)$ in (4.6) can be made small independent of $\xi \in K_0$. This finishes the proof of (5.5). □

6. The nonautonomous version. Finally, we come to the nonautonomous analogues of the main result. We must first reset our notation. Suppose $F: [0, \infty) \times \mathfrak{R}^n \rightrightarrows \mathfrak{R}^n$ is a multifunction with $[0, \infty) \times X \subseteq \operatorname{dom} F$, where $X \subseteq \mathfrak{R}^n$ is open. Consider the differential inclusion

$$\begin{aligned} &x(\cdot) \in AC[t_0, t_1] \\ &\dot{x}(t) \in F(t, x(t)) \text{ a.e. } t \in [t_0, t_1] \\ &x(t_0) = \xi, \end{aligned} \tag{6.1}$$

where $0 \leqq t_0 < t_1$. The reachable set is now defined by

$$R(t_0, t_1, \xi) = \{x(t_1): x(\cdot) \text{ satisfies (6.1)}\}.$$

Escape times now also depend on the initial time:

$$T_{X,t_0}(\xi) := \sup\left\{t_1: \operatorname{cl} \bigcup_{t_0 \leqq t \leqq t_1} R(t_0, t_1, \xi) \text{ is compact in } X\right\}.$$

Assume F has compact convex values and satisfies:

for $\xi \in X$, $t \rightrightarrows F(t, \xi)$ is continuous on $[0, \infty)$, and

for $T > 0$, $\xi \rightrightarrows f(t, \xi)$ is locally Lipschitz on X, independent of $t \in [0, T]$.

THEOREM 6.1. *Suppose F is as above. Then for all $\xi \in X$ and t_0 and t_1 satisfying $0 \leqq t_0 \leqq t_1 < T_{X,t_0}(\xi)$, we have*

$$R(t_0, t_1, \xi) = \lim_{N \to \infty} \left(\prod_{j=1}^{N} \left(I + \frac{(t_1 - t_0)}{N} F\left(t_0 + j\frac{t_1 - t_0}{N}\right)\right)\right)(\xi). \tag{6.2}$$

The proof of Theorem 6.1 involves a routine modification of the proof of Theorem 4.1, and therefore is omitted. It is also a straightforward matter to prove nonautonomous versions of Corollaries 5.1 and 5.2 and Proposition 5.3.

7. Two simple proofs. In this section, two well known properties of reachable sets are deduced effortlessly from the exponential formula. Again for notational simplicity, we return to the autonomous formulation (1.1). Throughout this section, X and F are as in Corollary 4.2.

COROLLARY 7.1 (cf. [1, p. 106]). *For all $\xi \in X$ and $0 < T < T_X(\xi)$, the reachable set $R^{(T)}(\xi)$ is connected.*

Proof. Observe two simple facts: (1) if G is a continuous multifunction with compact convex values, then the image under G of a connected set is connected. (2) the Hausdorff limit of connected sets is connected.

It follows from (1) that $(I + (T/N)F)^N(\xi)$ is connected for all N. From (2), connectedness is preserved in the limit. Hence, it is immediately deduced via Corollary 4.2 that $R^{(T)}(\xi)$ is connected. □

COROLLARY 7.2 (cf. [1, p. 120]). *Suppose $K_1 \subseteq K_2$ are two compact subsets of X, and suppose further that $T > 0$ and $\delta > 0$ so that $R^{(T)}(\xi) + \delta B \subseteq K_2$ for each $\xi \in K_1$ and $0 \leqq t \leqq T$. Let λ be a Lipschitz constant for F on K_2. Then the multifunction $\xi \rightrightarrows R^{(T)}(\xi)$ is Lipschitz of order $e^{\lambda T}$ on K_1.*

Proof. Again observe two simple facts: (1) If G_1 and G_2 are two multifunctions so that G_1 is Lipschitz of order λ_1 on K_1, and G_2 is Lipschitz of order λ_2 on $G_1(K_1)$, then $G_2 \circ G_1$ is Lipschitz of order $\lambda_1\lambda_2$ on K_1. (2) A pointwise limit of Lipschitz multifunctions is again Lipschitz with order less than or equal the lim sup of the orders of the sequence.

In view of Proposition 5.3, we have for each large N that $(I + (T/N)F)^j(\xi) \subseteq K_2$ for all $j = 0, 1, \cdots, N$ and $\xi \in K_1$. Since $(I + (T/N)F)$ is Lipschitz of order $1 + (T/N)\lambda$ on K_2, applying (1) N times gives us that $(I + (T/N)F)^N$ is Lipschitz of order $(1 + (T/N)\lambda)^N$ on K_1. Passing to the limit and using (2), we conclude $\xi \rightrightarrows R^{(T)}(\xi)$ is Lipschitz of order $e^{\lambda T}$ on K_1. □

8. Application to the Mayer optimal control problem. In this final section, we apply Proposition 5.3 with the purpose of characterizing the value function associated with the following control problem (which is formulated as a differential inclusion):

$$\inf f(x(T)) \text{ over } x(\cdot) \text{ satisfying (1.1).} \tag{8.1}$$

We assume: $X = \mathfrak{R}^n$ in (1.1); $F: \mathfrak{R}^n \rightrightarrows \mathfrak{R}^n$ is as in Corollary 4.2; all escape times are

greater than T; and $f:\mathfrak{R}^n \to \mathfrak{R}^1$ is locally Lipschitz. The assumption on the escape times holds for all T, for example, if F satisfies the growth condition $|F(x)| \leqq r_1|x| + r_2$.

The dynamic programming approach to providing optimality conditions to (8.1) consists of using properties of the associated value function $V(t, \eta)$, which is defined for $(t, \eta) \in [0, T] \times \mathfrak{R}^n$ as the optimal value of the problem

$$\begin{aligned} &\inf f(x(T)) \quad \text{over} \\ &x(\cdot) \in AC[t, T] \\ &\dot{x}(s) \in F(x(s)) \text{ a.e. } s \in [t, T] \\ &x(t) = \eta. \end{aligned} \tag{8.2}$$

Nonsmooth analysis has had a major role here, because in general $V(t, \eta)$ will not be differentiable at all points. Necessary and sufficient conditions for a trajectory to solve (8.1) in terms of a solution to a generalized Hamilton-Jacobi inequality is given by Clarke and Vinter [3]. Also see Vinter and Wolenski [10] where measurable time dependence of F is treated. The result of this section can be interpreted that if a uniformity condition is assumed in the limits, then the solution is unique (and, of course, equals the value function).

A successful approach to the uniqueness problem in Hamilton-Jacobi theory has been followed by Lions, Crandall, and others through the notion of a viscosity solution. This approach extends far beyond the relatively simple problem (8.1) we are treating here. Loosely speaking, in our analysis of (8.1), the viscosity supersolution corresponds to a generalized solution of the Hamilton-Jacobi inequality used in the optimality conditions. For our uniqueness result below, the concept of viscosity subsolution is replaced by the uniformity condition.

THEOREM 8.1. *The value function V is the unique function φ defined on $[0, T] \times \mathfrak{R}^n$ that satisfies*:

(i) *$\varphi(t, \eta)$ is locally Lipschitz in (t, η)*

(ii) *$\varphi(T, \eta) = f(\eta)$ for all $\eta \in \mathfrak{R}^n$*

(iii) *for each compact $K \subseteq \mathfrak{R}^n$ and $\varepsilon > 0$, there exists $\delta > 0$ so that*

$$\inf_{v \in F(\eta)} \frac{|\varphi(t+h, \eta+hv) - \varphi(t, \eta)|}{h} < \varepsilon \tag{8.3}$$

holds for all $\eta \in K$, $0 < h < \delta$, and $t \in [0, T-h]$.

Proof. It is well known that the value function is locally Lipschitz, and (ii) is trivial. We show (iii) holds for $\varphi = V$.

Let $K \subseteq \mathfrak{R}^n$ be compact and $\varepsilon > 0$. Then the set $M := \cup \{R^{(T)}(\eta): \eta \in K, 0 \leqq t \leqq T\}$ is bounded. This can be deduced by Gronwall's inequality (cf. [10, Lemma 3.1]) or it can also be seen as a simple consequence of Corollary 7.2. Let λ and l be Lipschitz constants for F and f, respectively, on $M + B$ (as usual, B denotes the closed unit ball).

Observe that $V(t, \eta) = \inf\{f(\gamma): \gamma \in R^{(T-t)}(\eta)\}$. Hence for all small h, we have for any $v \in \mathfrak{R}^n$:

$$\begin{aligned} \frac{1}{h}|V(t+h, \eta+hv) - V(t, \eta)| &\leqq \frac{l}{h} \operatorname{dist}_H (R^{(T-t-h)}(\eta+hv), R^{(T-t)}(\eta)) \\ &= \frac{l}{h} \operatorname{dist}_H (R^{(T-t-h)}(\eta+hv), R^{(T-t-h)}(R^{(h)}(\eta))) \\ &\leqq \frac{l}{h} e^{\lambda T} \operatorname{dist}_H (\{\eta+hv\}, R^{(h)}(\eta)). \end{aligned} \tag{8.4}$$

The last inequality is a simple consequence of Corollary 7.2. Hence taking the inf over $v \in F(\eta)$ in (8.4) gives

$$\inf_{v \in F(\eta)} \frac{|V(t+h, \eta+hv) - V(t, \eta)|}{h} \leqq l e^{\lambda T} \operatorname{dist}_H \left(F(\eta), \frac{R^{(h)}(\eta) - \eta}{h} \right). \tag{8.5}$$

From [11, Thm. 3.1(d)], the right side of (8.5) approaches zero as $h \downarrow 0$ uniformly over η in a compact set. Hence (8.3) holds for $\varphi = V$.

We now turn to the proof of the uniqueness assertion. Suppose $\varphi : [0, T] \times \mathfrak{R}^n \to \mathfrak{R}^1$ satisfies (i), (ii), and (iii). Then an immediate consequence of (iii) is that

$$\inf_{v \in F(\eta)} \liminf_{h \downarrow 0} \frac{\varphi(t+h, \eta+hv) - \varphi(t, \eta)}{h} = 0. \tag{8.6}$$

From [10, Thm. 2.3], it follows that $V \geqq \varphi$. (Actually for this estimate, φ needs only to satisfy (8.6) with the inequality "$\geqq$" replacing the equality.)

To obtain the reverse inequality $V \leqq \varphi$, we will employ Proposition 5.3. Let K be a compact set whose interior contains $\cup \{R^{(s)}(\eta) : T - t \leqq s \leqq T\} + B$. Let $0 < \varepsilon \leqq 1$, and fix $t < T$ and $\eta \in \mathfrak{R}^n$. For these choices of K and ε, choose δ so that (iii) holds, and choose N_0 large so that Proposition 5.3 holds. We may assume $(1/N_0) < \delta$.

We next define a discrete trajectory which is "ε-optimal." Set $h = (T - t/N_0)$ and $t_j = t + jh$, $j = 0, 1, \cdots, N$. Let $\eta_0 = \eta$, and suppose for some $k \leqq N - 1$, η_j is chosen for $j = 1, \cdots, k$ so that

$$\eta_j \in \eta_{j-1} + hF(\eta_{j-1}) \tag{8.7}$$

and

$$\frac{1}{h} |\varphi(t_j, \eta_j) - \varphi(t_{j-1}, \eta_{j-1})| < \varepsilon. \tag{8.8}$$

Note that (8.7) implies $\eta_j \in (I + hF)^j(\eta)$ for each $j = 1, \cdots, k$, and hence by Proposition 5.3, lies in K. By (iii) there exists $\eta_{k+1} \in \eta_k + hF(\eta_k)$ so that (8.8) holds for $j = k+1$. At stage N_0, a sequence $\{\eta_j\}_{j=0}^{N_0}$ is constructed satisfying (8.7) and (8.8) for all $0 \leqq j \leqq N_0$. Set $\bar{\eta} = \eta_{N_0}$. We have

$$\begin{aligned} |f(\bar{\eta}) - \varphi(t, \eta)| &= |\varphi(T, \bar{\eta}) - \varphi(t, \eta)| \quad \text{(by (ii))} \\ &\leqq h \sum_{j=1}^{N_0} \frac{1}{h} |\varphi(t_j, \eta_j) - \varphi(t_{j-1}, \eta_{j-1})| \\ &\leqq \varepsilon \quad \text{(by (8.8))}. \end{aligned} \tag{8.9}$$

We can now compare $\varphi(t, \eta)$ with $V(t, \eta)$:

$$\begin{aligned} V(t, \eta) &= \inf \{f(\gamma) : \gamma \in R^{(T-t)}(\eta)\} \\ &\leqq \inf \{f(\gamma) : \gamma \in (I + hF)^{N_0}(\eta)\} + l\varepsilon \quad \text{(by Proposition 5.3)} \\ &\leqq f(\bar{\eta}) + l\varepsilon \quad \text{(by construction of } \bar{\eta}) \\ &\leqq \varphi(t, \eta) + (1 + l)\varepsilon \quad \text{(by (8.9))}. \end{aligned}$$

Since l does not depend on ε, and ε is arbitrarily small, we conclude

$$V(t, \eta) \leqq \varphi(t, \eta).$$

This concludes the proof of Theorem 8.1. □

REFERENCES

[1] J. P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, Heidelberg, 1984.
[2] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley Interscience, New York, 1983.
[3] F. H. CLARKE AND R. B. VINTER, *Local optimality conditions and Lipschitzian solutions to the Hamilton-Jacobi equation*, SIAM J. Control Optim., 21 (1983), pp. 856-870.
[4] A. L. DONTCHEV AND E. M. FARKHI, *Error estimates for discretized differential inclusions*, Computing, 41 (1989), pp. 349-358.
[5] A. F. FILIPPOV, *Classical solutions of differential equations with multivalued right-hand side*, SIAM J. Control, 5 (1967), pp. 609-621.
[6] R. T. ROCKAFELLAR, *Convex algebra and duality in dynamic models of production*, in Mathematical Models in Economics, J. Loś and M. W. Loś, eds., Polish Scientific Publishers, Warszawa, 1974.
[7] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1974.
[8] K. TAUBERT, *Converging multistep methods for initial value problems involving multivalued maps*, Computing, 27 (1981), pp. 123-136.
[9] R. B. VINTER, *A characterization of the reachable set for nonlinear control systems*, SIAM J. Control Optim., 18 (1980), pp. 599-610.
[10] R. B. VINTER AND P. R. WOLENSKI, *Hamilton-Jacobi theory for optimal control problems with data measurable in time*, SIAM J. Control Optim., to appear.
[11] P. R. WOLENSKI, *A uniqueness theorem for differential inclusions*, J. Differential Equations, 83 (1990), to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 1162-1189, September 1990

008

REALIZATION OF AUTOREGRESSIVE EQUATIONS IN PENCIL AND DESCRIPTOR FORM*

M. KUIJPER† AND J. M. SCHUMACHER‡

Abstract. A linear system described by autoregressive equations with a given input/output structure cannot be transformed to standard state-space form if the implied input/output relation is nonproper. Instead, a realization in descriptor form must be used. In this paper, it is shown how to obtain minimal descriptor realizations from autoregressive equations without separating finite and infinite frequencies, and without going through a reduction process. External equivalence is used, so that even situations in which there is no transfer matrix can be considered. The approach is based on the so-called *pencil representation* of linear systems, and it is shown that there is a natural realization of autoregressive equations in pencil form. In this way, the link between the realization theories of Willems and Fuhrmann can also be clarified.

Key words. linear systems, autoregressive equations, descriptor form, pencil representation, realization, external equivalence

AMS(MOS) subject classifications. 93B15, 93B20

1. Introduction and preliminaries. In this paper, we study methods for obtaining state representations for linear systems given by higher-order equations in external variables, with special attention to the so-called "nonproper" situation. Suppose that relations between input variables u and output variables y are specified by equations of the form

$$R_1(\sigma)y + R_2(\sigma)u = 0 \tag{1.1}$$

where $R_1(\sigma)$ and $R_2(\sigma)$ are polynomial matrices, σ denotes differentiation or shift (depending on whether we work in continuous time or in discrete time), and y and u are functions of time. Here, as well as below, the time argument is suppressed to alleviate the notation. The argument σ will sometimes be replaced by λ or s; λ denotes a formal parameter, whereas s is used as a complex parameter and serves as default. Following the terminology of Willems [19], we will refer to (1.1) as a set of *autoregressive equations.* Inputs and outputs are jointly referred to as *external variables,* and (1.1) may be rewritten as

$$R(\sigma)w = 0 \tag{1.2}$$

where $R(s) = [R_1(s)\ \ R_2(s)]$ is sometimes called an *AR matrix,* and $w = [y^\top\ \ u^\top]^\top$ is the *vector of external variables.* Of course, it is also possible to take (1.2) as a starting point, without distinction between "inputs" and "outputs" in the external variables. The *behavior* defined by (1.2) is the set of all time functions w that satisfy (1.2). A behavior may also be specified by other means, for instance, by representations that involve *auxiliary* (*internal*) *variables,* such as the state representations to be defined below. Two representations will be said to be *externally equivalent* [18] if their induced behaviors are the same. In this paper, we will be looking for minimal representations under external equivalence. In comparison with the notion of transfer equivalence, which has been used more commonly in realization theory, external equivalence is both stronger and more general—more general, because transfer equivalence can be

* Received by the editors January 16, 1989; accepted for publication (in revised form) October 30, 1989.
† Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, the Netherlands.
‡ Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, the Netherlands and Department of Economics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, the Netherlands.

defined only for systems with a given input/output structure that is such that a transfer matrix exists, and stronger, because when both notions are applicable, external equivalence implies transfer equivalence but not the other way around. To avoid confusion, let us note that the notion of "external equivalence" as understood in [2] is different from the notion used here; for example, the systems $\dot{y} = \dot{u}$ and $y = u$ are equivalent in the sense of [2] but not in the sense of this paper.

The standard realization theory presupposes that the matrix $R_1(s)$ is square and nonsingular, and that $R_1^{-1}(s)R_2(s)$ is proper rational. Under these assumptions, it is well known that an equivalent representation can be found in the usual state-space form

$$\sigma x = Ax + Bu, \qquad y = Cx + Du. \tag{1.3}$$

A powerful and elegant method to obtain such a state-space realization was devised by Fuhrmann [5] who stated his result under transfer equivalence, and a similar procedure under external equivalence was given by Willems [19]. However, the standard assumptions mentioned above are not always satisfied. Examples of situations in which this occurs can be found, for instance, in circuit models [13], econometric models [11], and system inversion [7]. An often used modification of (1.3), that enables us to also cover these so-called *nonproper* situations, is the *descriptor* form [10]

$$\sigma Ex = Ax + Bu, \qquad y = Cx + Du \tag{1.4}$$

where the matrix E is not necessarily invertible. Algorithms to go from (1.1) to (1.4), which follow the line of [5], have been presented in [22] and [4]. Both papers work under transfer equivalence and so there is still the assumption that the matrix $R_1(s)$ is invertible. The realization procedure is then based on a decomposition of the transfer matrix $R_1^{-1}(s)R_2(s)$ into a strictly proper and a polynomial part. For the strictly proper part, a representation in standard state-space form is obtained by the usual means, and the polynomial part is realized in special descriptor form by using a modification of Fuhrmann's procedure; finally, the two realizations are put together again to create a representation in descriptor form.

One of the important uses of realization theory is the translation of properties of and statements about linear systems from polynomial terms to state-space terms and vice versa, as is extensively shown in [6]. The realization procedure for nonproper systems by cutting and pasting, as just described, is somewhat indirect, and is therefore less suitable for such translation purposes. In this paper, we will show how to obtain a realization in descriptor form without separation of finite and infinite frequencies. The realization will be obtained under external equivalence, and will be minimal in the appropriate sense. As an application, we will establish the relationships between basic indices associated with the representation (1.1) and with the representation (1.4). The realization procedure will be motivated along the lines of [19], and our discussion will also clarify the relationship between the realization algorithm in [19] and the one in [5].

The development below will be based on what we call the *pencil representation* of a linear system. This is a representation of the form

$$\sigma Gz = Fz, \qquad w = Hz \tag{1.5}$$

where w is a vector of external variables containing both inputs and outputs, and σ again denotes either differentiation or shift. A similar representation has been used before in [1], and pencil techniques in general are popular tools in numerical system theory (see, for instance, [16]). It may also be noted that the form (1.5) has been used for systems with partial differential equations in which control is exerted through the boundary conditions ("boundary control systems"; cf. [14]).

Formally, a pencil representation is given by a six-tuple $(Z, X, W; F, G, H)$ in which W is the space of external variables, Z is the space of internal variables, X is the equation space, F and G are linear mappings from Z to X, and H is a linear mapping from Z to W. We shall consider only pencil representations that are finite-dimensional in the sense that both $\dim Z$ and $\dim X$ are finite. Also, $\dim W$ will always be finite. Two pencil representations $(Z, X, W; F, G, H)$ and $(\tilde{Z}, \tilde{X}, W; \tilde{F}, \tilde{G}, \tilde{H})$ will be called *isomorphic* if there exist isomorphisms $S: Z \to \tilde{Z}$ and $T: X \to \tilde{X}$ such that $\tilde{G} = TGS^{-1}$, $\tilde{F} = TFS^{-1}$, and $\tilde{H} = HS^{-1}$. The *behavior* given by a pencil representation is the set of all w for which there exists a z such that (1.5) holds. (One has to select suitable function classes here; this will be discussed later.) A pencil representation is said to be *minimal* (under external equivalence) if both $\dim Z$ and $\dim X$ are minimal in the class of equivalent representations. Let us quickly review what can be inferred about minimality of pencil representations from the existing literature.

PROPOSITION 1.1. *A pencil representation $(Z, X, W; F, G, H)$ is minimal under external equivalence if and only if the following conditions hold:*

(i) *G is surjective;*

(ii) *$[G^\top H^\top]^\top$ is injective;*

(iii) *the matrix $[sG^\top - F^\top H^\top]^\top$ has full column rank for all $s \in \mathbb{C}$.*

Moreover, a minimal representation is unique up to isomorphism.

Proof. If G is not surjective in a representation of the form (1.5), then "Step One" of the realization algorithm in [15] may be used to find an equivalent representation with a smaller equation space X. So in every minimal representation the mapping G must be surjective. By a suitable choice of bases in X and Z, a matrix representation of G may then be given as $G = [I \ 0]$; with respect to these bases, write $F = [A \ B]$, and $H = [C' \ D']$. Writing z correspondingly as a vector with components ξ and η, the representation (1.5) takes the form

$$\sigma\xi = A\xi + B\eta \qquad w = C'\xi + D'\eta. \tag{1.6}$$

The variable η is known as the "driving variable" ([19]). It is known ([18, Thm. 4.5], [19, § 5], [15, Cor. 4.2]) that such a system is minimal if and only if $V^*(A, B, C', D') = \{0\}$ and D' is injective. The condition on V^* and the injectivity of D' together imply that the associated system pencil

$$\begin{pmatrix} sI - A & B \\ C' & D' \end{pmatrix} \tag{1.7}$$

has full column rank for all s (see [8, p. 544]), so that (iii) holds. Because D' is injective, the matrix

$$\begin{pmatrix} I & 0 \\ C' & D' \end{pmatrix}$$

is injective, also; this implies (ii). Conversely, if the conditions (i)–(iii) hold, then it follows from (ii) and (iii) that the system pencil has full column rank for all s, so that V^* in the equivalent state space form must be zero. The injectivity of D' in the equivalent state space form is immediate from (ii), by reversing the argument used above.

Now consider two minimal representations $(Z, X, W; F, G, H)$ and $(\tilde{Z}, \tilde{X}, W; \tilde{F}, \tilde{G}, \tilde{H})$ of the same system. As above, both representations can be rewritten in driving-variable form; the resulting state-space representations will be denoted by (A, B, C', D') and $(\tilde{A}, \tilde{B}, \tilde{C}', \tilde{D}')$, respectively. Because these are minimal representations of the same behavior, it follows from Theorem 7.1 in [18] that there exist invertible

mappings Q and R and a mapping F such that $\tilde{A} = Q(A+BF)Q^{-1}$, $\tilde{B} = QBR$, $\tilde{C}' = (C'+D'F)Q^{-1}$ and $\tilde{D}' = DR$. So we can write the following equations:

$$[I \quad 0] = Q[I \quad 0]\begin{pmatrix} Q^{-1} & 0 \\ FQ^{-1} & R \end{pmatrix} \tag{1.8}$$

$$[\tilde{A} \quad \tilde{B}] = Q[A \quad B]\begin{pmatrix} Q^{-1} & 0 \\ FQ^{-1} & R \end{pmatrix} \tag{1.9}$$

$$[\tilde{C}' \quad \tilde{D}'] = [C' \quad D']\begin{pmatrix} Q^{-1} & 0 \\ FQ^{-1} & R \end{pmatrix}. \tag{1.10}$$

This shows that the two given representations are isomorphic.

Remark 1.2. It is not hard to see that if (i) of the above proposition holds and the matrix $[sG^\top - F^\top \quad H^\top]^\top$ has full column rank (as a rational matrix), then condition (ii) holds if and only if $[sG^\top - F^\top \quad H^\top]^\top$ has no zeros at infinity. So, items (ii) and (iii) of the proposition may be replaced by the following two conditions:

(ii)′ *the matrix* $[sG^\top - F^\top \quad H^\top]^\top$ *has full column rank*;

(iii)′ *the matrix* $[sG^\top - F^\top \quad H^\top]^\top$ *has no zeros in the extended complex plane.*

2. Pencil representations from a given behavior: discrete time. In this section, we will discuss the pencil representation for systems that are given directly through their (discrete-time) behavior. Here our treatment is close to the development in [19]; however, we emphasize the pencil representation rather than the driving-variable representation, and we derive some results that do not depend on the assumption that the behavior is closed in the topology of pointwise convergence.

Following the definition in [19], a *linear, time-invariant, discrete-time behavior* is a shift-invariant subspace of the space $W^{\mathbb{Z}_+}$ of all functions from $\mathbb{Z}_+$ to a vector space $W \simeq \mathbb{R}^q$. The following mappings are defined on $W^{\mathbb{Z}_+}$: the *shift*

$$\sigma : (w_0, w_1, \cdots) \mapsto (w_1, w_2, \cdots), \tag{2.1}$$

the *forward shift*

$$\sigma^* : (w_0, w_1, \cdots) \mapsto (0, w_0, w_1, \cdots), \tag{2.2}$$

and the *evaluation mapping at time* 0

$$\chi : (w_0, w_1, \cdots) \mapsto w_0. \tag{2.3}$$

Now, let $\mathscr{B}$ be a given behavior. Following [19], we introduce the subspaces

$$\mathscr{B}^0 = \{w \in \mathscr{B} \mid (\sigma^*)^k w \in \mathscr{B} \ \forall k \geqq 0\} \tag{2.4}$$

and

$$\mathscr{B}^1 = \{w \in \mathscr{B}^0 \mid \chi w = 0\} \tag{2.5}$$

of $\mathscr{B}$. Intuitively, $\mathscr{B}^0$ contains the trajectories that start from the zero state; so the quotient space $\mathscr{B}/\mathscr{B}^0$ should be (isomorphic to) the state space. The quotient space $\mathscr{B}^0/\mathscr{B}^1$ describes the freedom that arises at each point in time because of the freedom we have in choosing a value of the input variable (or rather, a value of the "driving variable"). So, $\mathscr{B}^0/\mathscr{B}^1$ is the candidate for the space of driving variables. The following facts are trivially verified:

$$\sigma\mathscr{B}^1 \subset \mathscr{B}^0 \tag{2.6}$$

$$\mathscr{B}^1 \subset \ker \chi. \tag{2.7}$$

Because of (2.6), we can properly define a mapping $M_1: \mathscr{B}/\mathscr{B}^1 \to \mathscr{B}/\mathscr{B}^0$ by

$$M_1: w \bmod \mathscr{B}^1 \mapsto \sigma w \bmod \mathscr{B}^0. \tag{2.8}$$

Because of (2.7), there is also a mapping $M_2: \mathscr{B}/\mathscr{B}^1 \to W$ defined by

$$M_2: w \bmod \mathscr{B}^1 \mapsto \chi w. \tag{2.9}$$

Furthermore, we introduce the projection mapping $M_0: \mathscr{B}/\mathscr{B}^1 \to \mathscr{B}/\mathscr{B}^0$, defined simply by

$$M_0: w \bmod \mathscr{B}^1 \mapsto w \bmod \mathscr{B}^0. \tag{2.10}$$

If elements of $\mathscr{B}/\mathscr{B}^1$ are seen as "state + driving variable," then M_0 deletes the driving variable. The mappings M_0, M_1, and M_2 could also have been introduced by requiring that Fig. 1 below commutes, where π^0 denotes projection modulo $\mathscr{B}^0$ and π^1 projection modulo $\mathscr{B}^1$.

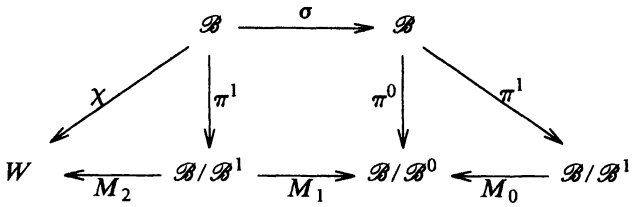


FIG. 1

The discrete-time behavior described by a *pencil representation* such as (1.5) will be denoted by $\mathscr{B}_p(Z, X, W; F, G, H)$. More explicitly,

$$\mathscr{B}_p(Z, X, W; F, G, H) = \{w: \mathbb{Z}_+ \to W \mid \exists z: \mathbb{Z}_+ \to \mathbb{Z} \text{ s.t. } \sigma Gz = Fz \text{ and } Hz = w\}. \tag{2.11}$$

We can now formulate the following proposition.

PROPOSITION 2.1. *For any linear, time-invariant, discrete-time behavior $\mathscr{B}$, we have*

$$\mathscr{B} \subset \mathscr{B}_p(\mathscr{B}/\mathscr{B}^1, \mathscr{B}/\mathscr{B}^0, W; M_1, M_0, M_2). \tag{2.12}$$

Proof. Take $w \in \mathscr{B}$. Define $z: \mathbb{Z}_+ \to \mathscr{B}/\mathscr{B}^1$ by

$$z_k = \pi^1 \sigma^k w. \tag{2.13}$$

From the definitions we easily verify that $\sigma M_0 z = M_1 z$ and that $M_2 z = w$. This proves that $w \in \mathscr{B}_p(\mathscr{B}/\mathscr{B}^1, \mathscr{B}/\mathscr{B}^0, W; M_1, M_0, M_2)$.

The *closure* of a behavior $\mathscr{B}$ (in the topology of pointwise convergence) will be denoted by $\mathscr{B}^{cl}$. A sequence w belongs to $\mathscr{B}^{cl}$ if and only if for every $k \geqq 0$ there exists a $\tilde{w} \in \mathscr{B}$ such that $w_j = \tilde{w}_j$ for all $0 \leqq j \leqq k$.

PROPOSITION 2.2. *For any linear, time-invariant, discrete-time behavior $\mathscr{B}$, we have*

$$\mathscr{B}^{cl} \supset \mathscr{B}_p(\mathscr{B}/\mathscr{B}^1, \mathscr{B}/\mathscr{B}^0, W; M_1, M_0, M_2). \tag{2.14}$$

Proof. Take $w \in \mathscr{B}_p(\mathscr{B}/\mathscr{B}^1, \mathscr{B}/\mathscr{B}^0, W; M_1, M_0, M_2)$, and let $z: \mathbb{Z}_+ \to \mathscr{B}/\mathscr{B}^1$ be such that $\sigma M_0 z = M_1 z$ and $M_2 z = w$. To show that $w \in \mathscr{B}^{cl}$, we will prove by induction that for every k there exists a $\tilde{w}^k \in \mathscr{B}$ such that $w_i = \tilde{w}_i^k$ for $0 \leqq i \leqq k$. First, let $\hat{w}^k \in \mathscr{B}$ be such that

$$z_k = \pi^1 \hat{w}^k. \tag{2.15}$$

Next, define $\tilde{w}^k$ by

$$\tilde{w}^k = (\hat{w}_0^0, \hat{w}_0^1, \cdots, \hat{w}_0^k, \hat{w}_1^k, \hat{w}_2^k, \cdots). \tag{2.16}$$

For $0 \leqq i \leqq k$, we have

$$w_i = M_2 z_i = M_2 \pi^1 \hat{w}^i = \chi \hat{w}^i = \hat{w}_0^i = \tilde{w}_i^k. \tag{2.17}$$

It remains to prove that $\tilde{w}^k \in \mathcal{B}$ for all k. For $k=0$, this is trivial since $\tilde{w}^0 = \hat{w}^0 \in \mathcal{B}$. Since

$$\begin{aligned}\tilde{w}^{k+1} - \tilde{w}^k &= (0, 0, \cdots, 0, \hat{w}_0^{k+1} - \hat{w}_1^k, \hat{w}_1^{k+1} - \hat{w}_2^k, \cdots) \\ &= (\sigma^*)^k(\hat{w}^{k+1} - \sigma\hat{w}^k),\end{aligned} \tag{2.18}$$

the proof will follow by induction if we can show that $\hat{w}^{k+1} - \sigma\hat{w}^k \in \mathcal{B}^0$ for all k. But this follows from

$$\pi^0\hat{w}^{k+1} = M_0\pi^1\hat{w}^{k+1} = M_0 z_{k+1} = M_1 z_k = M_1\pi^1\hat{w}^k = \pi^0\sigma\hat{w}^k. \tag{2.19}$$

COROLLARY 2.3 [19]. *If* $\mathcal{B} = \mathcal{B}^{\mathrm{cl}}$, *then* $\mathcal{B}_{\mathrm{p}}(\mathcal{B}/\mathcal{B}^1, \mathcal{B}/\mathcal{B}^0, W; M_1, M_0, M_2) = \mathcal{B}$.

The above corollary states that every closed, linear, time-invariant behavior admits a pencil representation. Moreover, as shown in [19, Thm. 9], the spaces $\mathcal{B}/\mathcal{B}^1$ and $\mathcal{B}/\mathcal{B}^0$ that appear in the representation $\mathcal{B}_{\mathrm{p}}(\mathcal{B}/\mathcal{B}^1, \mathcal{B}/\mathcal{B}^0, W; M_1, M_0, M_2)$ are *finite-dimensional.* For completeness, we will offer a proof of this fact which we think is more straightforward than the two proofs that were already given for essentially the same fact in [19]. Some notation will be needed. Let $[w]_k$ denote the *k-truncation* of an element w of $W^{\mathbb{Z}_+}$; if

$$w = (w_0, w_1, \cdots, w_k, w_{k+1}, \cdots), \tag{2.20}$$

then

$$[w]_k = (w_0, w_1, \cdots, w_k). \tag{2.21}$$

For subspaces $\mathcal{B}$ of $W^{\mathbb{Z}_+}$, write

$$\mathcal{B}_k = \{[w]_k \mid w \in \mathcal{B}\}. \tag{2.22}$$

Define a sequence of subspaces of W by

$$W_k^0(\mathcal{B}) = \{w \in W \mid (0, 0, \cdots, 0, w) \in \mathcal{B}_k\}. \tag{2.23}$$

We shall let $\mathcal{B}$ be a fixed linear time-invariant behavior, and write W_k^0 rather than $W_k^0(\mathcal{B})$. It is immediate from $\sigma\mathcal{B} \subset \mathcal{B}$ that $W_{k+1}^0 \subset W_k^0$ for all k. Because W is finite-dimensional, the sequence of subspaces $W_0^0 \supset W_1^0 \supset \cdots$ must reach a limit after a finite number of steps; the limit subspace will be denoted by W^0. We now prove the following lemma.

LEMMA 2.4. *Suppose that* $\mathcal{B}$ *is closed. Let* k_0 *be such that* $W_{k_0}^0 = W^0$, *and let* $\Phi : \mathcal{B} \to \mathcal{B}_{k_0}$ *denote the mapping* $w \mapsto [w]_{k_0}$. *Under these conditions, we have*

$$\ker \Phi \subset \mathcal{B}^0. \tag{2.24}$$

Proof. Since $\mathcal{B}^0$ is by definition the largest σ^*-invariant subspace of $\mathcal{B}$, it suffices to show that $\ker\Phi$ is σ^*-invariant. Take $w \in \ker\Phi$; we want to show that also $\sigma^* w \in \ker\Phi$, which will follow if we can prove that $\sigma^* w \in \mathcal{B}$. For this, it is sufficient to show that

$$[\sigma^* w]_j \in \mathcal{B}_j \quad \forall j \geqq 0, \tag{2.25}$$

by the closedness of $\mathcal{B}$. For $0 \leqq j \leqq k_0 + 1$, $[\sigma^* w]_j = 0$ and so the condition (2.25) is certainly satisfied. To proceed by induction, suppose that $[\sigma^* w]_i \in \mathcal{B}_i$ for some $i \geqq k_0 + 1$. Let $\tilde{w} \in \mathcal{B}$ be such that $[\tilde{w}]_i = [\sigma^* w]_i$. We then have $[w - \sigma\tilde{w}]_{i-1} = 0$, and therefore,

$$w_i - \tilde{w}_{i+1} \in W_i^0 = W_{i+1}^0. \tag{2.26}$$

From (2.26) and the fact that $[\sigma^* w - \tilde{w}]_i = 0$, it follows that

$$[\sigma^* w - \tilde{w}]_{i+1} \in \mathcal{B}_{i+1}. \tag{2.27}$$

Since $[\tilde{w}]_{i+1}$ obviously belongs to $\mathscr{B}_{i+1}$, we may conclude that $[\sigma^* w]_{i+1} \in \mathscr{B}_{i+1}$, which is what we wanted to prove.

Remark 2.5. From the lemma, we easily derive that W^0, the limit of the sequence in (2.23), is equal to $\chi\mathscr{B}^0$.

PROPOSITION 2.6. *If a linear, time-invariant behavior $\mathscr{B}$ is closed, then $\mathscr{B}/\mathscr{B}^0$ is finite-dimensional.*

Proof. By the lemma, we have

$$\dim \mathscr{B}/\mathscr{B}^0 \leqq \dim \mathscr{B}/\ker \Phi = \dim \operatorname{im} \Phi \leqq \dim W^{k_0+1} = q(k_0+1). \tag{2.28}$$

It is not hard to show directly that the pencil representation obtained above is, in fact, minimal.

LEMMA 2.7. *If $(Z, X, W; F, G, H)$ is a pencil representation of the linear, time-invariant behavior $\mathscr{B}$, then*

$$\dim X \geqq \dim \mathscr{B}/\mathscr{B}^0 \tag{2.29}$$

and

$$\dim Z \geqq \dim \mathscr{B}/\mathscr{B}^1. \tag{2.30}$$

Proof. Introduce the behavior of the auxiliary variables

$$\mathscr{Z} = \{z : \mathbb{Z}_+ \mapsto Z \mid \sigma G z = F z\}. \tag{2.31}$$

By definition of a pencil representation, we have

$$H\mathscr{Z} = \mathscr{B}. \tag{2.32}$$

In analogy with $\mathscr{B}^0$, we also introduce

$$\mathscr{Z}^0 = \{z \in \mathscr{Z} \mid (\sigma^*)^k z \in \mathscr{Z} \ \forall k \geqq 0\}. \tag{2.33}$$

Obviously, we have

$$H\mathscr{Z}^0 \subset \mathscr{B}^0. \tag{2.34}$$

It is easily verified that, in fact,

$$\mathscr{Z}^0 = \{z \in \mathscr{Z} \mid G z_0 = 0\}, \tag{2.35}$$

which shows that $\mathscr{Z}^0$ is the kernel of the mapping which assigns the element Gz_0 of X to a given $z \in \mathscr{Z}$. As a consequence, we get

$$\dim (\mathscr{Z}/\mathscr{Z}^0) \leqq \dim X. \tag{2.36}$$

Because of (2.34), we can unambiguously define a mapping $\Psi : \mathscr{Z}/\mathscr{Z}^0 \to \mathscr{B}/\mathscr{B}^0$ by

$$\Psi : z \bmod \mathscr{Z}^0 \mapsto Hz \bmod \mathscr{B}^0. \tag{2.37}$$

Moreover, (2.32) shows that this map is surjective. Therefore,

$$\dim \mathscr{B}/\mathscr{B}^0 \leqq \dim \mathscr{Z}/\mathscr{Z}^0 \leqq \dim X. \tag{2.38}$$

For the proof of the second inequality, we introduce

$$\mathscr{Z}^1 = \{z \in \mathscr{Z}^0 \mid z_0 = 0\} = \{z \in \mathscr{Z} \mid z_0 = 0\} \tag{2.39}$$

and proceed analogously, noting that $H\mathscr{Z}^1 \subset \mathscr{B}^1$ and that $\dim(\mathscr{Z}/\mathscr{Z}^1) \leqq \dim Z$.

We summarize the main results in the following theorem.

THEOREM 2.8. *Let $\mathscr{B}$ be a closed, linear, time-invariant, discrete-time behavior. Then a finite-dimensional minimal pencil representation of $\mathscr{B}$ is given by $(\mathscr{B}/\mathscr{B}^1, \mathscr{B}/\mathscr{B}^0, W; M_1, M_0, M_2)$, where $\mathscr{B}^0$ and $\mathscr{B}^1$ are defined by* (2.4) *and* (2.5), *respectively, and the mappings M_0, M_1, and M_2 are defined by requiring that Fig.* 1 *commutes.*

A behavior $\mathscr{B}$ will rarely be given "as such," and consequently the construction of a pencil representation as given above is mainly of theoretical value. Two important ways of prescribing a behavior are the following:

- by *data*: $\mathscr{B}$ is determined as the smallest closed, linear, shift-invariant subspace of $W^{\mathbb{Z}_+}$ that contains a given (finite) set of trajectories. This leads to realization procedures involving generalizations of the Hankel matrix: see [20] and, for the case of approximate modeling, [21].
- by *equations*: $\mathscr{B}$ is determined as the set of all trajectories that satisfy a certain set of differential or difference equations. For the purpose of describing a closed, linear, time-invariant behavior, such equations may always be rewritten in the form $R(\sigma)w = 0$, where $R(s)$ is a polynomial matrix [18, Prop. 3.3].

We shall be concerned with the second option in this paper. In the next section, we shall consider systems given by a set of equations $R(\sigma)w = 0$, and we shall construct a pencil representation by expressing the spaces $\mathscr{B}/\mathscr{B}^0$, etc. in terms of the polynomial matrix $R(s)$.

3. Pencil representations from autoregressive equations: discrete time. Let a behavior be given by

$$R(\sigma)w = 0 \tag{3.1}$$

where $R(s)$ is a polynomial matrix of size $k \times q$, and σ denotes the shift. We shall continue to work in discrete time in order to employ the results of the previous section to give a representation in pencil form for the behavior described by (3.1). Similar results can be obtained for systems in continuous time, but these require a different proof technique and will be handled in the next section.

It will be convenient to use an alternative notation for time series, one that is more adapted to the description in terms of a polynomial matrix. Via the correspondence

$$(w_0, w_1, \cdots) \leftrightarrow w_0\lambda^{-1} + w_1\lambda^{-2} + \cdots, \tag{3.2}$$

we can identify $W^{\mathbb{Z}_+}$ with the set of formal power series (with vanishing constant term) in the parameter λ^{-1}. This set, to be denoted by ΩW, is a subset of the set ΛW of formal Laurent series around infinity in λ, of which a typical element is

$$w_{-i-1}\lambda^{i} + w_{-i}\lambda^{i-1} + \cdots + w_{-1} + w_0\lambda^{-1} + w_1\lambda^{-2} + \cdots.$$

The natural projection of ΛW onto ΩW, effected by "deleting the polynomial part," will be denoted by π_-. Elements of ΩW will be written as $w(\lambda)$ or sometimes also simply as w.

The action of the shift σ on $W^{\mathbb{Z}_+}$ corresponds on ΩW to multiplication by λ followed by projection:

$$\sigma w \leftrightarrow \pi_-(\lambda w(\lambda)). \tag{3.3}$$

Consequently, the behavior $\mathscr{B}$ given by (3.1) is represented in ΩW by the set X^R that is defined by

$$X^R = \{w \in \Omega W \mid \pi_-(R(\lambda)w(\lambda)) = 0\}. \tag{3.4}$$

The right shift σ^* is represented in ΩW by multiplication by λ^{-1}. Therefore, $\mathscr{B}^0$ corresponds to the subspace N^R defined by

$$N^R = \{w \in \Omega W \mid \pi_-(\lambda^{-k}R(\lambda)w(\lambda)) = 0 \ \forall k \geqq 0\} = \{w \in \Omega W \mid R(\lambda)w(\lambda) = 0\}. \tag{3.5}$$

Finally, $\mathscr{B}^1$ is equal to $\sigma^*\mathscr{B}^0$, which corresponds to $\lambda^{-1}N^R$.

The quotient space $\mathscr{B}/\mathscr{B}^0$, which plays a role in the pencil representation of the previous section as the space in which the dynamic equation "takes place," is represented as X^R/N^R. We can consider mutiplication by $R(\lambda)$ as a mapping from X^R to $\mathbb{R}^k[\lambda]$, the set of polynomials with coefficients in $\mathbb{R}^k$. The space N^R is then precisely the kernel of this mapping, which suggests replacing the quotient space X^R/N^R by the isomorphic space

$$X_R = \{p(\lambda) \in \mathbb{R}^k[\lambda] \mid \exists w(\lambda) \in \Omega W \text{ s.t. } R(\lambda)w(\lambda) = p(\lambda)\}. \tag{3.6}$$

The isomorphism is given, of course, by the mapping M_R defined as follows:

$$M_R : w(\lambda) \bmod N^R \mapsto R(\lambda)w(\lambda) \qquad (w(\lambda) \in X^R). \tag{3.7}$$

With some of the notation used in Fig. 1 unchanged, we now introduce the mappings F, G, and H by requiring that Fig. 2 below commutes. We then obtain the following theorem.

THEOREM 3.1. *The behavior given by* (3.1) *is equal to* $\mathscr{B}_p(X^R/\lambda^{-1}N^R, X_R, W; F, G, H)$; *and this pencil representation is minimal.*

Proof. Apart from changes of notation, all we did was replace the representation derived in the previous section by an isomorphic one. The result is therefore immediate from Theorem 2.8.

Bases for the vector spaces X_R and $X^R/\lambda^{-1}N^R$ may be found by taking $R(s)$ to row reduced form, and then concrete matrix representations for the mappings F, G, and H can be obtained. This is worked out in § 8.

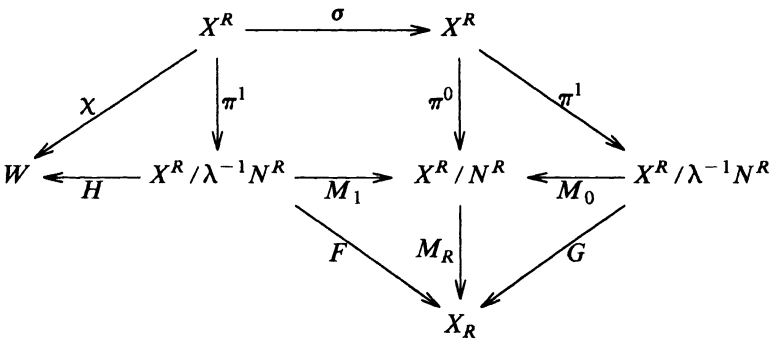


FIG. 2

4. Pencil representations from autoregressive equations: continuous time. In the discrete-time context, many system properties are conveniently expressed in terms of the behavior itself, and we have used this fact extensively in the previous sections to prove properties of representations; for instance, equivalence between AR and pencil representations could be proved by reducing both to their associated behaviors. For systems in continuous time, however, the representation of a behavior in terms of itself is much less manageable, and we are forced to work with representations in terms of equations. The formal definition of a continuous-time behavior requires the specification of a function class to which the trajectories should belong. We will denote by $\mathscr{F}$ the function class to which the (components of the) trajectories of the external variables belong; the class from which the components of the trajectories of internal (auxiliary) variables are taken will be denoted by $\mathscr{D}$. We will assume that $\mathscr{D}$ is a linear function space that is closed under differentiation and that contains $\mathscr{F}$; differential equations will always be considered in the sense of $\mathscr{D}$. All properties used below will be valid when $\mathscr{F} = \mathscr{D} = C^\infty(\mathbb{R})$ (see for instance [15]), but other choices are also possible—however, we shall not go into the axiomatics here. Confer also the discussion in [3, Chaps. 4, 5]. The development below may also be applied to systems in discrete time, although the approach of the preceding two sections would seem to be preferable for its intuitive appeal.

We begin by noting some facts concerning the elimination of auxiliary variables. To interpret the statements in the lemma below, it is useful to remember that with any behavior $\mathscr{B}$ admitting an AR representation we can associate a subspace of the rational vector space $W(\lambda)$ of rational W-valued functions in the formal parameter λ. Indeed, if $R(s)$ is an AR matrix for the given behavior, then $R(\lambda)$ can be viewed as a mapping between rational vector spaces, and its kernel is easily seen to be independent of the choice of the representation. So $\ker R(\lambda)$ is uniquely determined by the behavior. In the interpretation of the previous section, $\ker R(\lambda)$ is just the linear span (over $\mathbb{R}(\lambda)$) of the elements of $\mathscr{B}$. In particular, $\dim \ker R(\lambda)$ is the number of inputs in any standard state space description of $\mathscr{B}$.

LEMMA 4.1. *Consider a behavior $\mathscr{B}$ given by the equations*

$$P(\sigma)\xi = 0 \tag{4.1}$$

$$w = Q(\sigma)\xi \tag{4.2}$$

where $P(s)$ and $Q(s)$ are polynomial matrices, and ξ contains auxiliary variables. Denote by q the number of rows of $Q(s)$, by n the number of rows of $P(s)$, and by r the rank of $[P^\top(s)\ \ Q^\top(s)]^\top$. It is always possible to find polynomial matrices $V(s)$ and $R(s)$ such that

(i) *$V(s)$ has size $(n+q-r)\times n$, $R(s)$ has size $(n+q-r)\times q$;*

(ii) *$V(s)$ and $R(s)$ are left coprime, i.e., the matrix $[V(s)\ \ R(s)]$ has full row rank for all $s \in \mathbb{C}$;*

(iii) *$V(s)P(s)+R(s)Q(s)=0$.*

If $V(s)$ and $R(s)$ satisfy these properties, then an AR *description of the behavior defined by* (4.1)-(4.2) *is*

$$R(\sigma)w = 0, \tag{4.3}$$

and the following relation holds, where all matrices are interpreted as matrices over the field of rational functions:

$$\ker R(\lambda) = Q(\lambda)[\ker P(\lambda)]. \tag{4.4}$$

In particular, we have

$$\dim \ker R(\lambda) = \operatorname{rank}\begin{pmatrix} P(\lambda) \\ Q(\lambda) \end{pmatrix} - \operatorname{rank} P(\lambda). \tag{4.5}$$

Proof. For instance by reduction to Hermite form [8, p. 375] we can find a unimodular matrix $U(s)$ of size $(n+q)\times(n+q)$ such that

$$\begin{pmatrix} U_{11}(s) & U_{12}(s) \\ U_{21}(s) & U_{22}(s) \end{pmatrix}\begin{pmatrix} P(s) \\ Q(s) \end{pmatrix} = \begin{pmatrix} T(s) \\ 0 \end{pmatrix} \tag{4.6}$$

where $T(s)$ has full row rank. Clearly then, the number of rows of $T(s)$ must be equal to r, and so the dimensions of $U_{21}(s)$ and $U_{22}(s)$ are $(n+q-r)\times n$ and $(n+q-r)\times q$, respectively. It is easily verified also that conditions (ii) and (iii) above are satisfied by taking $V(s) = U_{21}(s)$ and $R(s) = U_{22}(s)$.

Suppose now that $V(s)$ and $R(s)$ satisfy conditions (i)-(iii). We can then find polynomial matrices $U_1(s)$ and $U_2(s)$ such that the matrix

$$\begin{pmatrix} U_1(s) & U_2(s) \\ V(s) & R(s) \end{pmatrix}$$

is unimodular. If we write $T(s) = U_1(s)P(s) + U_2(s)Q(s)$, then we obviously have

$$\begin{pmatrix} U_1(s) & U_2(s) \\ V(s) & R(s) \end{pmatrix}\begin{pmatrix} P(s) \\ Q(s) \end{pmatrix} = \begin{pmatrix} T(s) \\ 0 \end{pmatrix}. \tag{4.7}$$

Moreover, $T(s)$ must be of full row rank, since its number of rows is equal to the rank of $[P^{\top}(s)\ \ Q^{\top}(s)]^{\top}$. This implies that $R(s)$ is an AR matrix for the behavior given by (4.1)-(4.2) (see [15, Cor. 2.3]). The formula (4.4) is obtained by interpreting (4.7) as a rational matrix equation and using straightforward linear algebra, and (4.5) is an immediate consequence. This completes the proof of the lemma.

In the discrete-time context, we used quotients of sequence spaces to construct the vector spaces that are needed in a pencil representation. It should be noted that the end result would have been the same if we would have replaced the sequence spaces by corresponding spaces of rational vector functions; in particular, the space $W(\lambda)$ of rational functions with values in W may be substituted for ΛW, and $\lambda^{-1}W[[\lambda^{-1}]]$ (the space of strictly proper rational W-valued functions) for ΩW. For continuous-time systems, the use of sequence spaces is less natural, and we shall use the rational setting. This will also facilitate comparison with the results of Fuhrmann (see, e.g., [6]). The symbol π_- will be used now for the natural projection of $X(\lambda)$ (where X is any vector space) onto $\lambda^{-1}X[[\lambda^{-1}]]$. For an element $w(\lambda)$ of $\lambda^{-1}W[[\lambda^{-1}]]$, the value of $sw(s)$ at infinity will be denoted by w_{-1} in accordance with the notation of [6], rather than by w_0 as would be suggested by (3.2).

The next theorem is the main result of this section. Essentially, it shows how to solve the equations that we obtain by requiring that Fig. 2 commutes.

THEOREM 4.2. *Let a system be given in* AR *form* (1.2), *with* $R(s)\in\mathbb{R}^{k\times q}(s)$ *of full row rank. Consider the following spaces of rational vector functions in a formal parameter* λ:

$$X^R = \{w(\lambda)\in\lambda^{-1}W[[\lambda^{-1}]] \mid \pi_- R(\lambda)w(\lambda) = 0\}, \tag{4.8}$$

$$X_R = \{p(\lambda)\in\mathbb{R}^k[\lambda] \mid \exists w(\lambda)\in\lambda^{-1}W[[\lambda^{-1}]] \text{ s.t. } p(\lambda) = R(\lambda)w(\lambda)\}, \tag{4.9}$$

$$N^R = \{w(\lambda)\in\lambda^{-1}W[[\lambda^{-1}]] \mid R(\lambda)w(\lambda) = 0\}. \tag{4.10}$$

The following mappings (G *and* F *from* $X^R/\lambda^{-1}N^R$ *to* X_R, H *from* $X^R/\lambda^{-1}N^R$ *to* W) *are well defined*:

$$G: w(\lambda) \bmod \lambda^{-1}N^R \mapsto R(\lambda)w(\lambda), \tag{4.11}$$

$$F: w(\lambda) \bmod \lambda^{-1}N^R \mapsto R(\lambda)\pi_-(\lambda w(\lambda)), \tag{4.12}$$

$$H: w(\lambda) \bmod \lambda^{-1}N^R \mapsto w_{-1}. \tag{4.13}$$

With these definitions, ($X^R/\lambda^{-1}N^R$, X_R, W; F, G, H) *is a minimal pencil representation of the behavior given by* (1.2).

Proof. It is easily verified that the mappings F, G, and H are indeed well-defined. Because $\lambda^{-1}N^R$ is contained in N^R, it is obvious from the definition (4.11) that G is surjective. If, for some $w(\lambda) \in X^R$, both $w_{-1} = 0$ and $R(\lambda)w(\lambda) = 0$, then $\lambda w(\lambda)$ belongs to N^R so $w(\lambda)$ belongs to $\lambda^{-1}N^R$. This shows that the mapping $[G^\top \ H^\top]^\top$ is injective. Furthermore, suppose that $s \in \mathbb{C}$ and $w(\lambda) \in X^R$ are such that we have

$$sR(\lambda)w(\lambda) - R(\lambda)\pi_-(\lambda w(\lambda)) = 0 \tag{4.14}$$

$$w_{-1} = 0. \tag{4.15}$$

Because of (4.15), $\pi_-(\lambda w(\lambda))$ is equal to $\lambda w(\lambda)$, and (4.14) may be rewritten as

$$(s-\lambda)R(\lambda)w(\lambda) = 0. \tag{4.16}$$

Of course, this implies that $R(\lambda)w(\lambda) = 0$. Because we also have (4.15), it follows that $w(\lambda) \in \lambda^{-1}N^R$. By the definitions, this shows that $[sG^\top - F^\top \ H^\top]^\top$ is injective for all $s \in \mathbb{C}$. By the criterion given in Proposition 1.1, we have now shown that the pencil representation given by F, G, and H is minimal.

We still must show that this pencil representation describes the same behavior as the given AR representation. For this purpose, we use the preceding lemma. Let n denote the dimension of X_R and write r for the dimension of $X^R/\lambda^{-1}N^R$; then r is also the rank of $[sG^\top - F^\top \ H^\top]^\top$, since we have shown that this matrix has full column rank. Because G is surjective and $\ker G = N^R/\lambda^{-1}N^R$, we can write

$$r - n = \dim \ker G = \dim N^R/\lambda^{-1}N^R = \dim \ker R(\lambda) = q - k \tag{4.17}$$

since $R(s)$ was assumed to be of full row rank. So, we have $k = n + q - r$, and $R(s)$ has the size required in Lemma 4.1. It remains to find a polynomial matrix $V(s)$ of size $k \times n$ such that conditions (ii) and (iii) of that lemma are satisfied.

We claim that such a polynomial mapping is given by the "evaluation map" which replaces the formal parameter λ by the complex number s:

$$V(s): X_R \ni p(\lambda) \mapsto p(s) \in \mathbb{C}^k. \tag{4.18}$$

This map is polynomial because X_R consists of polynomial vectors; this is evident when we write a matrix representation of $V(s)$. To verify that condition (ii) holds, we compute, for $w(\lambda) \in X^R$:

$$\begin{aligned} V(s)(sG - F)w(\lambda) &= V(s)[sR(\lambda)w(\lambda) - R(\lambda)(\lambda w(\lambda) - w_{-1})] \\ &= sR(s)w(s) - R(s)(sw(s) - w_{-1}) \\ &= R(s)w_{-1} = R(s)Hw(\lambda). \end{aligned} \tag{4.19}$$

Finally, we must show that $V(s)$ and $R(s)$ are left coprime. For this purpose, it suffices to produce polynomial mappings $Q_1(s)$ and $Q_2(s)$ such that

$$V(s)Q_1(s) + R(s)Q_2(s) = I. \tag{4.20}$$

By assumption, $R(s)$ has full row rank, so it has a rational right inverse, say $T(s)$. We split $T(s)$ into a polynomial and a strictly proper part, denoted, respectively, by $T_+(s)$ and $T_-(s)$. Obviously, we have

$$R(s)T_-(s) = I - R(s)T_+(s), \tag{4.21}$$

where the right-hand side is polynomial. It follows that the columns of $R(\lambda)T_-(\lambda)$ belong to X_R. Consequently, there exists a constant matrix Q_1 such that

$$R(s)T_-(s) = V(s)Q_1. \tag{4.22}$$

Writing $T_+(s)$ as $Q_2(s)$, we get

$$V(s)Q_1 + R(s)Q_2(s) = R(s)T_-(s) + R(s)T_+(s) = R(s)T(s) = I. \tag{4.23}$$

5. Realization with a causal input/output structure. In the realization procedure of the previous section, we could replace the quotient space X^R/N^R by the space of polynomials X_R, because we had a natural isomorphism available between these two spaces, given essentially by multiplication by $R(\lambda)$. The other space that we used, $X^R/\lambda^{-1}N^R$, is isomorphic to the direct sum $X_R \oplus W^0$, where W^0 is the subspace of W defined by

$$W^0 = \{w \in W \mid \exists w(\lambda) \in N^R \text{ s.t. } w = w_{-1}\}. \tag{5.1}$$

(In other words, we have $W^0 = HN^R$, in full analogy with the discrete-time case—see Remark 2.5.) Indeed, the following holds:

$$X^R/\lambda^{-1}N^R \simeq X^R/N^R \oplus N^R/\lambda^{-1}N^R \simeq X_R \oplus W^0. \tag{5.2}$$

Unfortunately, the first isomorphism in the formula above must be established by selecting a complement to $N^R/\lambda^{-1}N^R$ in $X^R/\lambda^{-1}N^R$, and so we do not have a *natural* isomorphism available. This is also reflected in the nonuniqueness of "driving-variable" representations as described in [18, Thm. 7.1]. It should be noted that the space W^0 itself is canonically given (i.e., it is an invariant under external equivalence), and this space will play an important role below.

Now, suppose that we add more structure by dividing the external variables into *inputs* and *outputs*. Such a division is given by a decomposition of the external variable space W as the direct sum of two subspaces Y and U, corresponding to a splitting of the defining AR matrix $R(s)$ as

$$R(s) = [R_1(s) \quad R_2(s)]. \tag{5.3}$$

The projection onto U along Y will be denoted by π_U, the complementary projection by π_Y. We shall first consider the "causal" situation as described in the following lemma, which is a formalization of remarks in [18, § 6]. General input/output structures will be discussed in the next section.

LEMMA 5.1. *With the notations introduced above, the following statements are equivalent*:

(i) $R_1(s)$ *is invertible as a rational matrix, and* $R_1^{-1}(s)R_2(s)$ *is proper rational*;

(ii) *the projection* π_U, *taken as a mapping from* W^0 *to* U, *is an isomorphism*;

(iii) *there exists a mapping* $D: U \mapsto Y$ *such that*

$$W^0 = \left\{ \begin{pmatrix} Du \\ u \end{pmatrix} \middle| u \in U \right\} \tag{5.4}$$

where the vector notation is adapted to the decomposition of W *as* $Y \oplus U$;

(iv) *Y is a complement of W^0 in W.*

Proof. The equivalence between statements (ii), (iii), and (iv) is a matter of straightforward linear algebra. To prove that (i) implies (iii), define

$$D=[-R_1^{-1}(s)R_2(s)]_{s=\infty}. \tag{5.5}$$

Take $w\in W^0$, and let $w(\lambda)\in N^R$ be such that $w_{-1}=w$. From $R(\lambda)w(\lambda)=0$, we have

$$\pi_Y w(\lambda)+R_1^{-1}(\lambda)R_2(\lambda)\pi_U w(\lambda)=0, \tag{5.6}$$

and this implies

$$\pi_Y w_{-1}=D\pi_U w_{-1}. \tag{5.7}$$

Conversely, suppose that $w\in W$ is of the form

$$w=\begin{pmatrix} Du \\ u \end{pmatrix}. \tag{5.8}$$

Define $w(\lambda)$ by

$$w(\lambda)=\lambda^{-1}\begin{pmatrix} R_1^{-1}(\lambda)R_2(\lambda)u \\ u \end{pmatrix}; \tag{5.9}$$

then $w(\lambda)\in N^R$ and $w_{-1}=w$, so that $w\in W^0$.

Now, assume that (ii)-(iv) hold. Let $N(\lambda)$ be a basis matrix for the rational vector space $\ker R(\lambda)$; we may assume that $N(\lambda)$ is proper rational, and that its leading coefficient matrix $N_0=[N(s)]_{s=\infty}$ has full column rank. (To see this, note that by reducing $R(\lambda)$ to row reduced form one actually writes $R(\lambda)=[S(\lambda)\ 0]B(\lambda)$ where $S(\lambda)$ is a nonsingular polynomial matrix, and $B(\lambda)$ is bicausal. One may then take $N(\lambda)=B^{-1}(\lambda)[0\ I]^{\top}$.) Under these conditions, N_0 is a basis matrix for W^0 and it follows that $\dim U=\dim W^0=q-k$ where k is the number of rows of $R(\lambda)$. So, $\dim Y=k$ and it is seen that the matrix $R_1(\lambda)$ is square. To prove that $R_1(\lambda)$ is invertible, suppose that $R_1(\lambda)y(\lambda)=0$ for some $y(\lambda)\in Y(\lambda)$ not equal to zero. It is no restriction of the generality to assume that $y(\lambda)$ is strictly proper with a nonzero leading term y_{-1}; but then the vector $[y_{-1}^{\top}\ 0]^{\top}$ belongs to $Y\cap W^0$ and so should be zero according to (iv). Finally, note that by definition we have

$$R_1(\lambda)\pi_Y N(\lambda)+R_2(\lambda)\pi_U N(\lambda)=0. \tag{5.10}$$

Moreover, the rational matrix $\pi_U N(\lambda)$ is proper with an invertible leading coefficient matrix, as is seen from (ii), and this implies that

$$R_1^{-1}(\lambda)R_2(\lambda)=-\pi_Y N(\lambda)(\pi_U N(\lambda))^{-1} \tag{5.11}$$

is proper rational. This completes the proof of the lemma.

In the remainder of this section, we assume that $R_1(s)$ in (5.3) is invertible, and that $R_1^{-1}(s)R_2(s)$ is proper rational. To construct the parameters in a standard state-space representation of the behavior given by $R(s)$, define a mapping Φ from $X^R/\lambda^{-1}N^R$ to $X_R\oplus U$ by

$$\Phi: w(\lambda)\bmod \lambda^{-1}N^R\mapsto\begin{pmatrix} R(\lambda)w(\lambda) \\ \pi_U w_{-1} \end{pmatrix} \tag{5.12}$$

(it is easily seen that this is well-defined). To prove that Φ is injective, let $w(\lambda)\in X^R$ be such that $R(\lambda)w(\lambda)=0$ and $\pi_U w_{-1}=0$. For such a $w(\lambda)$, we get $w(\lambda)\in N^R$ so $w_{-1}\in W^0$. The condition $\pi_U w_{-1}=0$ implies $w_{-1}\in Y$, so that $w_{-1}\in Y\cap W^0=\{0\}$, which proves that $w(\lambda)\in\lambda^{-1}N^R$. This shows that Φ is injective; the fact that Φ is actually

an isomorphism then follows easily by a dimension argument. Using the obvious facts $[I\ \ 0]\Phi = G$ and $[0\ \ I]\Phi = \pi_U H$, we can now write down the diagram below which we use to define the mappings A, B, C, and D that will appear in an input/state/output representation of the given behavior.

We can give more explicit expressions for the four mappings defined by requiring that Fig. 3 commutes. Note that $R_1^{-1}(s)p(s)$ is strictly proper if $p(\lambda) \in X_R$; indeed, suppose that $p(s) = R_1(s)\pi_Y w(s) + R_2(s)\pi_U w(s)$ for $w(\lambda) \in X^R$, then

$$R_1^{-1}(s)p(s) = \pi_Y w(s) + R_1^{-1}(s)R_2(s)\pi_U(s)w(s) \tag{5.13}$$

and this is obviously strictly proper. With this information, it is easily seen that the inverse of the isomorphism Φ may be given as follows:

$$\Phi^{-1}: X_R \oplus U \ni \begin{pmatrix} p(\lambda) \\ u \end{pmatrix} \mapsto \begin{pmatrix} R_1^{-1}(\lambda)p(\lambda) - \lambda^{-1}R_1^{-1}(\lambda)R_2(\lambda)u \\ \lambda^{-1}u \end{pmatrix} \bmod \lambda^{-1}N^R. \tag{5.14}$$

The mapping $[A\ \ B]$ can now be computed as $M_R M_1 \Phi^{-1}$. Explicitly, this gives:

$$\begin{aligned} [A\ \ B]\begin{pmatrix} p(\lambda) \\ u \end{pmatrix} &= [R_1(\lambda)\ \ R_2(\lambda)]\pi_-\lambda \begin{pmatrix} R_1^{-1}(\lambda)p(\lambda) - \lambda^{-1}R_1^{-1}(\lambda)R_2(\lambda)u \\ \lambda^{-1}u \end{pmatrix} \\ &= R_1(\lambda)\pi_-\lambda R_1^{-1}(\lambda)(p(\lambda) - R_2(\lambda)u) + R_2(\lambda)\pi_- u \\ &= \pi_{R_1}\lambda p(\lambda) - \pi_{R_1}R_2(\lambda)u, \end{aligned} \tag{5.15}$$

where the notation π_{R_1} is used, following [6], for the projection on X_R given by

$$\pi_{R_1}: p(\lambda) \mapsto R_1(\lambda)\pi_- R_1^{-1}(\lambda)p(\lambda). \tag{5.16}$$

In particular, we find

$$A: p(\lambda) \mapsto \pi_{R_1}\lambda p(\lambda) = \lambda p(\lambda) - R_1(\lambda)[R_1^{-1}(\lambda)p(\lambda)]_{-1} \tag{5.17}$$

and

$$B: u \mapsto -\pi_{R_1}R_2(\lambda)u. \tag{5.18}$$

The expression for B may also be written in a different way if we introduce a constant matrix D_∞ by

$$D_\infty = [R_1^{-1}(s)R_2(s)]_{s=\infty}; \tag{5.19}$$

namely,

$$B: u \mapsto -R_2(\lambda)u + R_1(\lambda)D_\infty u. \tag{5.20}$$

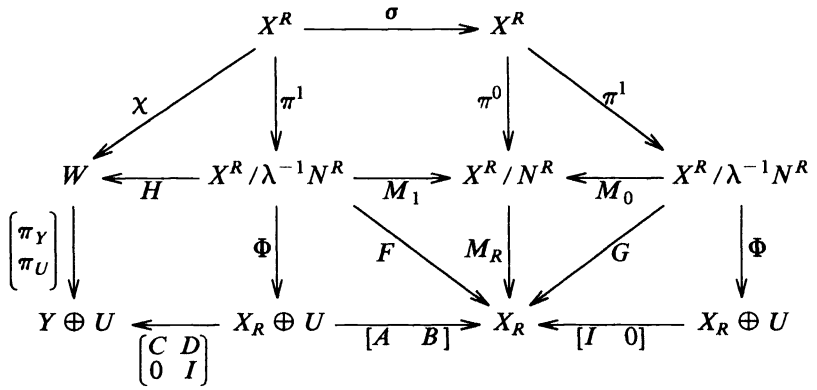


FIG. 3

Quite similarly, we obtain explicit expressions for the mappings C and D from the formula $[C\ D] = \pi_Y H \Phi^{-1}$. We find

$$C : p(\lambda) \mapsto [R_1^{-1}(\lambda) p(\lambda)]_{-1} \tag{5.21}$$

and

$$D : u \mapsto -D_\infty u. \tag{5.22}$$

So, in this way we recover Fuhrmann's realization of a transfer matrix $-R_1^{-1}(s)R_2(s)$ in left matrix fractional representation. Notice that actually we proved more: it is known from Fuhrmann's work that the realization is minimal under transfer equivalence if and only if the fractional representation is coprime, whereas we have shown here that the realization is *always* minimal under *external* equivalence. The condition for minimality under transfer equivalence can be derived from this.

It is also possible to set up diagrams to define single mappings from the quadruple (A, B, C, D). For instance, by transforming Fig. 3 we obtain Fig. 4, which can be used to define the mapping A. This clearly displays A as a version of the shift.

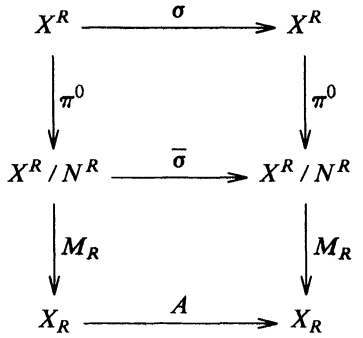


FIG. 4

6. Realization with a general input/output structure. In the case where we have given a not necessarily causal input/output description, our aim is to obtain a representation in descriptor form. To arrive at this representation, it turns out to be advantageous to use the pencil form as an intermediate step; the descriptor form can be derived from the pencil form in a straightforward way, as will now be shown.

Let a pencil representation $(Z, X, W; F, G, H)$ be given, along with a decomposition $W = Y \oplus U$ and associated projections π_Y and π_U. Decompose the internal variable space Z as $Z_0 \oplus Z_1 \oplus Z_2$ where $Z_1 = \ker G \cap \ker \pi_U H$, and $Z_1 \oplus Z_2 = \ker G$. Accordingly, write

$$G = [G_0 \quad 0 \quad 0], \qquad F = [F_0 \quad F_1 \quad F_2], \tag{6.1}$$

$$\pi_Y H = [H_{00} \quad H_{01} \quad H_{02}], \qquad \pi_U H = [H_{u0} \quad 0 \quad H_{u2}]. \tag{6.2}$$

The matrix H_{u2} has full column rank, and by renumbering the u-variables if necessary, we can write

$$H_{u0} = \begin{pmatrix} H_{10} \\ H_{20} \end{pmatrix}, \qquad H_{u2} = \begin{pmatrix} H_{12} \\ H_{22} \end{pmatrix} \tag{6.3}$$

where H_{22} is invertible (or empty, if $\ker G \subset \ker \pi_U H$). The system equations take the form (in obvious notation):

$$\sigma G_0 z_0 = F_0 z_0 + F_1 z_1 + F_2 z_2, \tag{6.4}$$

$$y = H_{00} z_0 + H_{01} z_1 + H_{02} z_2, \tag{6.5}$$

$$u_1 = H_{10} z_0 + H_{12} z_2, \tag{6.6}$$

$$u_2 = H_{20} z_0 + H_{22} z_2. \tag{6.7}$$

We can now solve for z_2 and obtain a description in descriptor form

$$\sigma E z = A z + B u, \tag{6.8}$$

$$y = C z + D u \tag{6.9}$$

where the parameters are defined as follows:

$$E = \begin{pmatrix} G_0 & 0 \\ 0 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} F_0 - F_2 H_{22}^{-1} H_{20} & F_1 \\ H_{10} - H_{12} H_{22}^{-1} H_{20} & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & F_2 H_{22}^{-1} \\ -I & H_{12} H_{22}^{-1} \end{pmatrix}, \tag{6.10}$$

$$C = [H_{00} - H_{02} H_{22}^{-1} H_{20} \quad H_{01}], \qquad D = [0 \quad H_{02} H_{22}^{-1}].$$

Remark 6.1. The essence of the above construction is that as many z-variables as possible are replaced by u-variables. If this is not considered important, then, of course a simpler construction is possible: just write

$$\sigma \begin{pmatrix} G \\ 0 \end{pmatrix} z = \begin{pmatrix} F \\ \pi_U H \end{pmatrix} z + \begin{pmatrix} 0 \\ -I \end{pmatrix} u, \tag{6.11}$$

$$y = \pi_Y H z. \tag{6.12}$$

This simple solution will in general produce a nonminimal descriptor representation even if one starts with a minimal pencil representation. The more elaborate construction above behaves nicely with respect to minimality properties, as shown below and as further detailed in [9].

The following lemma, which will be needed below, also sheds some light on the role of the u_2-variables. Recall that, in the construction above, these variables serve to parametrize the subspace $\pi_U H[\ker G]$ of W.

LEMMA 6.2. *Consider a pencil representation* (1.5) *and an equivalent* AR *representation* (1.2); *assume that G is surjective and that $[G^\top \; H^\top]^\top$ is injective. Let the subspace W^0 of W be defined by* (5.1). *We then have*

$$W^0 = H[\ker G]. \tag{6.13}$$

Proof. It follows from Lemma 4.1 that a rational vector $w(\lambda)$ belongs to $\ker R(\lambda)$ if and only if there exists a rational vector $z(\lambda)$ such that

$$\begin{pmatrix} 0 \\ w(\lambda) \end{pmatrix} = \begin{pmatrix} \lambda G - F \\ H \end{pmatrix} z(\lambda). \tag{6.14}$$

Now assume that $w(\lambda)$ is strictly proper; because $[G^\top \; H^\top]^\top$ is injective, it then follows that $z(\lambda)$ is also strictly proper, and that its leading coefficient z_{-1} satisfies $Gz_{-1} = 0$. Moreover, we have $w_{-1} = Hz_{-1}$. If follows that $W^0 \subset H[\ker G]$. Now, it has already been shown in the proof of Lemma 5.1 that $\dim W^0 = \dim \ker R(\lambda)$. Moreover, using (4.5) and the assumptions, we obtain

$$\dim \ker R(\lambda) = \operatorname{rank} \begin{pmatrix} \lambda G - F \\ H \end{pmatrix} - \operatorname{rank} (\lambda G - F) = \dim \ker (\lambda G - F) = \dim \ker G \tag{6.15}$$

so that $\dim \ker G = \dim W^0$. Since $\dim \ker G = \dim H[\ker G]$ because $[G^\top \; H^\top]^\top$ is injective, this leads to the desired conclusion.

Note that for *minimal* pencil representations, this characterization of W^0 in pencil terms can also be derived from the realization in § 4.

7. Indices and minimality. In this section, we will discuss the minimality of descriptor representations. While for standard state space systems there is only one index that plays a role to determine the minimality (viz., the dimension of the state space), there are three such indices for descriptor systems: the *rank* of E, the *column defect* of E ($\dim \ker E$ = the number of columns minus the rank), and the *row defect* of E ($\operatorname{codim} \operatorname{im} E$ = the number of rows minus the rank). A *minimal* descriptor representation is, by definition, one in which each of these three indices is minimal within the set of descriptor representations for a given behavior. Note that, with this definition, even the existence of a minimal representation is not trivial. Our strategy will be to establish first lower bounds for each of the three indices separately, and to show next that these minima can be achieved simultaneously. The fact that this is possible also shows that, by minimizing the three indices above, one automatically minimizes the number of descriptor variables (= the number of columns of E = rank + column defect) and the number of equations (= the number of rows of E = rank + row defect).

PROPOSITION 7.1. *Let an input/output behavior be given by autoregressive equations*

$$[R_1(\sigma) \quad R_2(\sigma)]\begin{pmatrix} y \\ u \end{pmatrix} = 0. \tag{7.1}$$

Write n for the sum of the minimal row indices of $R(s)$ (stated in other terms, n is the maximal degree of the full-size minors of $R(s)$). Suppose that a descriptor representation of the behavior determined by (7.1) *is given by*

$$\sigma E\xi = A\xi + Bu, \tag{7.2}$$

$$y = C\xi + Du. \tag{7.3}$$

Under these conditions, the rank of E is at least equal to n.

Proof. By a suitable choice of coordinates and introduction of new variables, the descriptor equations (7.2)-(7.3) may be written as follows:

$$\sigma\xi_1 = A_{11}\xi_1 + A_{12}\xi_2 + B_1\eta, \tag{7.4}$$

$$0 = A_{21}\xi_1 + A_{22}\xi_2 + B_2\eta, \tag{7.5}$$

$$\begin{pmatrix} y \\ u \end{pmatrix} = \begin{pmatrix} C & D \\ 0 & I \end{pmatrix}\begin{pmatrix} \xi \\ \eta \end{pmatrix}. \tag{7.6}$$

The algorithm of [15] may be used to reduce this to state-space (driving-variable) form; the dimension of the state space will be at most equal to the length of the vector ξ_1, which in turn is equal to the rank of E. On the other hand, it is well known (see [19, Thm. 6]) that the dimension of the state space must be at least equal to the sum of the minimal row indices of $R(s)$. The stated result follows.

The following two lemmas show that both observability at infinity and reachability at infinity (see for instance [12]) are necessary conditions for minimality of descriptor representations under external equivalence. This is unlike the situation for the finite modes, where minimality under external equivalence requires observability but not controllability [18, Cor. 4.7].

LEMMA 7.2. *A necessary condition for* (7.2)-(7.3) *to be a minimal descriptor representation is that the matrix $[E^\top \; C^\top]^\top$ is injective.*

Proof. Suppose that the condition of the lemma is not satisfied, so that ker E and ker C have a nontrivial intersection. By a suitable choice of coordinates, we may then write

$$E = [E_1 \quad 0], \qquad C = [C_1 \quad 0] \tag{7.7}$$

where the number of the columns in the zero matrices is equal to dim (ker $E \cap$ ker C). The equations (7.2)-(7.3) will then appear in the form

$$\sigma E_1 \xi_1 = A_{11}\xi_1 + A_{12}\xi_2 + Bu, \tag{7.8}$$

$$y = C_1\xi_1 + Du. \tag{7.9}$$

Denote the "equation space" (the space into which E maps) by X_e. Let X_e' and $T: X_e \to X_e'$ be such that T is surjective and satisfies ker $T =$ im A_{12}. The equations (7.8)-(7.9) are equivalent to

$$\sigma TE_1\xi_1 = TA_{11}\xi_1 + TBu, \tag{7.10}$$

$$y = C_1\xi_1 + Du. \tag{7.11}$$

We want to show that this system precedes the original system in the partial ordering determined by the three indices (rank, column defect, row defect) introduced above. That is, we want to show that the following inequalities hold, with strict inequality in at least one case:

$$\text{rank } TE_1 \leqq \text{rank } E, \tag{7.12}$$

$$\dim \ker TE_1 \leqq \dim \ker E, \tag{7.13}$$

$$\text{codim im } TE_1 \leqq \text{codim im } E. \tag{7.14}$$

As to (7.12), we have

$$\begin{aligned} \dim \text{im } TE_1 &= \dim \text{im } E_1 - \dim(\ker T \cap \text{im } E_1) \\ &\leqq \dim \text{im } E_1 = \dim \text{im } E \end{aligned} \tag{7.15}$$

with equality if and only if

$$\text{im } A_{12} \cap \text{im } E_1 = \{0\}. \tag{7.16}$$

We next consider (7.13):

$$\begin{aligned} \dim \ker TE_1 &= \dim \ker E_1 + \dim (\text{im } E_1 \cap \text{im } A_{12}) \\ &\leqq \dim \ker E_1 + \dim (\ker E \cap \ker C) = \dim \ker E \end{aligned} \tag{7.17}$$

where we used the fact that the number of columns of A_{12} is equal to dim (ker $E \cap$ ker C). Here, equality holds if and only if A_{12} has full column rank and

$$\text{im } A_{12} \subset \text{im } E_1. \tag{7.18}$$

Finally, we verify (7.14):

$$\text{codim im } TE_1 = \text{codim } T[\text{im } E_1] \leqq \text{codim im } E_1 = \text{codim im } E \tag{7.19}$$

with equality if and only if ker $T \subset$ im E_1, that is, if and only if (7.18) holds. (Here we use the following easily verified fact from linear algebra: if A is a surjective mapping from a space X to a space Y, and X_0 is a subspace of X, then codim $AX_0 \leqq$ codim X_0; equality holds if and only if ker $A \subset X_0$.)

Now, assume that equality would hold in all three cases. The matrix A_{12} should then have full column rank, so that the rank of A should equal the number of columns of A_{12}, which in its turn is equal to dim (ker $C \cap$ ker E). On the other hand, it follows from (7.16) and (7.18) that $A_{12} = 0$, so that it would follow that dim (ker $C \cap$ ker E) $= 0$, which contradicts our assumption that the subspaces ker C and ker E intersect non-trivially. This completes the proof.

LEMMA 7.3. *A necessary condition for* (7.2)-(7.3) *to be a minimal descriptor representation is that the matrix* $[E \ B]$ *is surjective.*

Proof. The proof is quite similar to the proof of the previous lemma, and we will not work out all details. Suppose that $[E \ B]$ is not surjective; then, by a suitable choice of coordinates, we can write

$$E = \begin{pmatrix} E_1 \\ 0 \end{pmatrix}, \qquad B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix} \tag{7.20}$$

where $[E_1 \ B_1]$ is surjective, and the number of zero rows is equal to codim $[E \ B]$. With this choice of coordinates, the equations (7.2)-(7.3) can be written as follows:

$$\sigma E_1 \xi = A_1 \xi + B_1 u, \tag{7.21}$$

$$0 = A_2 \xi, \tag{7.22}$$

$$y = C\xi + Du. \tag{7.23}$$

Let S be an injective mapping such that im $S =$ ker A_2. The above equations are equivalent to:

$$\sigma E_1 S\tilde{\xi} = A_1 S\tilde{\xi} + B_1 u, \tag{7.24}$$

$$y = CS\tilde{\xi} + Du. \tag{7.25}$$

To prove the lemma, we need to show that the following three inequalities hold, with strict inequality in at least one case:

$$\text{codim im } E_1 S \leqq \text{codim im } E, \tag{7.26}$$

$$\dim \ker E_1 S \leqq \dim \ker E, \tag{7.27}$$

$$\text{rank } E_1 S \leqq \text{rank } E. \tag{7.28}$$

This proof can be conducted as above (or the statement can be derived from the one in the previous lemma by duality).

PROPOSITION 7.4. *Let* (7.2)-(7.3) *be a descriptor representation for the behavior described by* (7.1), *and define* W^0 *as in* (5.1). *Under these conditions, the following inequalities hold*:

$$\dim \ker E \geqq \dim (Y \cap W^0), \tag{7.29}$$

$$\text{codim im } E \geqq \text{codim } (Y + W^0). \tag{7.30}$$

Proof. It follows from the lemmas we just proved that we may suppose that the matrix $[E^\top \ C^\top]^\top$ is injective and that the matrix $[E \ B]$ is surjective. Note that the descriptor equations (7.2)-(7.3) may also be written in the following form:

$$[\sigma E - A \quad -B] \begin{pmatrix} \xi \\ \eta \end{pmatrix} = 0, \tag{7.31}$$

$$\begin{pmatrix} y \\ u \end{pmatrix} = \begin{pmatrix} C & D \\ 0 & I \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix}. \tag{7.32}$$

Take $w \in W^0$; then there exists a proper rational W-valued function $w(\lambda)$ satisfying $w_0 = w$ and $R(\lambda)w(\lambda) = 0$. By Lemma 4.1 above, there must exist rational vector functions $\xi(\lambda)$ and $\eta(\lambda)$ such that

$$[\lambda E - A \quad -B]\begin{pmatrix}\xi(\lambda)\\ \eta(\lambda)\end{pmatrix} = 0 \tag{7.33}$$

$$w(\lambda) = \begin{pmatrix}C & D\\ 0 & I\end{pmatrix}\begin{pmatrix}\xi(\lambda)\\ \eta(\lambda)\end{pmatrix}. \tag{7.34}$$

These equations may also be written as follows:

$$\begin{pmatrix}\lambda E - A\\ C\end{pmatrix}\xi(\lambda) = \begin{pmatrix}0 & B\\ I & -D\end{pmatrix}w(\lambda). \tag{7.35}$$

Since the right-hand side in this equation is proper rational and because $[E^\top \quad C^\top]^\top$ is injective, $\xi(\lambda)$ must also be proper rational. Moreover, the constant term in the power series development of $\xi(\lambda)$ must satisfy $E\xi_0 = 0$. Now, suppose that $w \in Y \cap W^0$. Then, again from (7.35), it follows that $w = C\xi_0$; so $w \in C[\ker E]$. Therefore,

$$\dim (Y \cap W^0) \leqq \dim C\,[\ker E] = \dim \ker E. \tag{7.36}$$

For the proof of the second part, we note that it suffices to show that

$$\{u \in U \mid Bu \in \operatorname{im} E\} \subset \pi_U W^0. \tag{7.37}$$

Indeed, we easily verify that codim $\pi_U W^0$ (with $\pi_U W^0$ considered as a subspace of U) is equal to codim $(Y + W^0)$, and we can apply the following rule which holds generally for mappings A between vector spaces X and Y: codim $A^{-1}Y_0 \leqq$ codim Y_0 (Y_0 a subspace of Y). To show (7.37), let $u \in U$ be such that $Bu \in \operatorname{im} E$. The desired conclusion will follow if we can exhibit proper rational functions $\xi(\lambda)$ and $u(\lambda)$ such that $u_0 = u$ and

$$(\lambda E - A)\xi(\lambda) = Bu(\lambda). \tag{7.38}$$

If we define $y(\lambda) = C\xi(\lambda) + Du(\lambda)$, then $y(\lambda)$ is proper rational and

$$\begin{pmatrix}\lambda E - A\\ C\end{pmatrix}\xi(\lambda) = \begin{pmatrix}0 & B\\ I & -D\end{pmatrix}\begin{pmatrix}y(\lambda)\\ u(\lambda)\end{pmatrix} \tag{7.39}$$

so that

$$u = \pi_U \begin{pmatrix}y_0\\ u_0\end{pmatrix} \in \pi_U W^0. \tag{7.40}$$

Writing $u(\lambda) = u_0 + \eta(\lambda)$, we see that it will be sufficient to find a *strictly* proper solution $[\xi(\lambda)^\top \quad \eta(\lambda)^\top]^\top$ of the equation

$$[\lambda E - A \quad -B]\begin{pmatrix}\xi(\lambda)\\ \eta(\lambda)\end{pmatrix} = Bu. \tag{7.41}$$

Equivalently, we are looking for a *proper* solution of the same equation with Bu replaced by λBu. It follows from Theorem 6.3.12 in [8] that such a solution does indeed exist.

Remark 7.5. Actually, it is not difficult to display an explicit strictly proper solution to (7.41), if we rewrite this equation by a change of variables as

$$\begin{pmatrix} \lambda I - A_{11} & -A_{12} & -B_{11} & -B_{12} \\ -A_{21} & -A_{22} & I & -B_{22} \end{pmatrix} \begin{pmatrix} \xi_1(\lambda) \\ \xi_2(\lambda) \\ \eta_1(\lambda) \\ \eta_2(\lambda) \end{pmatrix} = \begin{pmatrix} x_0 \\ 0 \end{pmatrix}. \tag{7.42}$$

(The identity matrix in the (2, 3) position is allowed by the assumption that $[E\ B]$ is surjective.) A strictly proper solution is

$$\begin{pmatrix} \xi_1(\lambda) \\ \xi_2(\lambda) \\ \eta_1(\lambda) \\ \eta_2(\lambda) \end{pmatrix} = \begin{pmatrix} (\lambda I - A_{11} - B_{11}A_{21})^{-1}x_0 \\ 0 \\ A_{21}(\lambda I - A_{11} - B_{11}A_{21})^{-1}x_0 \\ 0 \end{pmatrix}, \tag{7.43}$$

as can be verified immediately.

THEOREM 7.6. *Let an input/output behavior be given by autoregressive equations* (7.1). *Denote the sum of the minimal indices of* $R(s)$ *by* n, *and define* W^0 *by* (5.1). *There exists an externally equivalent descriptor representation* (7.2)-(7.3) *satisfying the following requirements*:

$$\text{rank } E = n, \tag{7.44}$$

$$\dim \ker E = \dim (Y \cap W^0), \tag{7.45}$$

$$\text{codim im } E = \text{codim } (Y + W^0). \tag{7.46}$$

Moreover, a descriptor representation of the behavior given by (7.1) *is minimal if and only if the above three equalities hold.*

Proof. In view of the previous results in this section, it only remains to show that a descriptor representation satisfying (7.44)-(7.46) exists. We claim that the representation obtained in the previous section satisfies all requirements, supposing that this representation is formed from a minimal pencil representation (see Proposition 1.1). Using the notation of § 6, we have indeed:

$$\text{rank } E = \dim Z_0 = \dim \text{im } G = \dim X_R = n \tag{7.47}$$

$$\begin{aligned} Y \cap W^0 &= \ker \pi_U \cap H[\ker G] \\ &= \{w \in W \mid \exists z \in Z\colon Gz = 0,\ w = Hz,\ \pi_U w = 0\} \\ &= H[\ker G \cap \ker \pi_U H] = HZ_1 \end{aligned} \tag{7.48}$$

$$\dim \ker E = \dim Z_1 = \dim (\ker \pi_U H \cap \ker G) = \dim (Y \cap W^0) \tag{7.49}$$

(because $\ker G \cap \ker H = \{0\}$, so that the restriction of H to Z_1 is injective), and

$$\text{codim im } E = \dim U_1 = \text{codim } \pi_U W^0 = \text{codim } (Y + W^0). \tag{7.50}$$

Remark 7.7. By unimodular operations, we can take the given polynomial matrix $R(s)$ to row proper form (see [8, p. 386]); so, we may assume that $R(s)$ is row proper to start with. This means that we can write

$$R(s) = \Delta(s)B(s), \tag{7.51}$$

where $B(s)$ is right bicausal, and

$$\Delta(s) = \text{diag } (s^{\kappa_1}, \cdots, s^{\kappa_k}). \tag{7.52}$$

It is not difficult to verify that the subspace W^0 is characterized in these terms as

$$W^0 = \ker B(\infty). \tag{7.53}$$

Note that $B(\infty)$ is nothing but the "leading row coefficient matrix" of $R(s)$. The partitioning of $R(s)$ as $[R_1(s)\ \ R_2(s)]$ induces a similar partitioning of $B(\infty)$:

$$B(\infty)=[B_1(\infty)\quad B_2(\infty)]. \tag{7.54}$$

Using standard manipulations, we find the following expressions for $\dim(Y\cap W^0)$ and $\operatorname{codim}(Y+W^0)$:

$$\dim(Y\cap W^0)=\dim\ker B_1(\infty) \tag{7.55}$$

$$\operatorname{codim}(Y+W^0)=\operatorname{codim}\operatorname{im} B_1(\infty). \tag{7.56}$$

So, we have easy criteria for minimality of descriptor representations of a behavior given by a row proper AR matrix: the rank of E should be equal to the sum of the row indices, and the row and column defects of E should be equal to the corresponding indices of $B_1(\infty)$. It also follows that E in a minimal descriptor representation will be square if and only if $B_1(\infty)$ is square; this happens if and only if $R_1(s)$ is square, that is if the number of y-variables is equal to the number of independent equations in an AR representation.

8. Computation. In this section, we will show how to obtain concrete matrix representations in pencil form and in descriptor form, starting from autoregressive equations determined by a $k\times q$ polynomial matrix $R(s)$ of full row rank. For this purpose, we shall construct specific bases for the spaces that appear in the abstract realization of § 4. In the procedure below, the transformation from pencil to descriptor form is not a straightforward implementation of the abstract procedure given in § 6; one reason for this is that, in the abstract version, the crucial subspace W^0 appears as the *image* of a certain mapping, whereas in the computation below it appears as a *kernel*. This leads to a different (dual) method of selecting the u_2-variables.

The first step is to take the given polynomial matrix $R(s)$ to row proper form [8, p. 386]. To alleviate the notation, the resulting equivalent AR matrix will still be denoted by $R(s)$. So we have

$$R(s)=\Delta(s)B(s) \tag{8.1}$$

where $B(s)$ is right bicausal, and

$$\Delta(s)=\operatorname{diag}(s^{\kappa_1},\cdots,s^{\kappa_k}). \tag{8.2}$$

Now, let $\tilde{B}(s)$ be any matrix such that $\hat{B}(s)=[B^\top(s)\ \ \tilde{B}^\top(s)]^\top$ is bicausal. It will be discussed later how to make a suitable choice for $\tilde{B}(s)$. We can write $R(s)=[\Delta(s)\ \ 0]\hat{B}(s)$, and it is seen from this that a basis for $X^R/\lambda^{-1}N^R$ is given by the equivalence classes modulo $\lambda^{-1}N^R$ of the columns of the following matrix of size $q\times(n+q-k)$:

$$\hat{B}^{-1}(\lambda)\begin{bmatrix} \lambda^{-1}\cdots\lambda^{-\kappa_1} & & & & & \\ & \cdots & & & & \\ & & \lambda^{-1}\cdots\lambda^{-\kappa_k} & & & \\ & & & \lambda^{-1} & & \\ & & & & \ddots & \\ & & & & & \lambda^{-1} \end{bmatrix}. \tag{8.3}$$

A basis matrix for X_R is given by the following matrix of size $k\times n$:

$$\begin{bmatrix} \lambda^{\kappa_1-1}\cdots\lambda & 1 & 0 & & \\ 0 & & \lambda^{\kappa_2-1} & \cdots & \\ \vdots & & 0 & & \\ 0 & & \vdots & & \lambda^{\kappa_k-1}\cdots\lambda\ \ 1 \end{bmatrix}. \tag{8.4}$$

With respect to these bases, we now compute the matrix forms of F, G, and H. It is easily seen that G will take the form $[I\ \ 0]$. Because $\hat{B}(\lambda)$ is bicausal, the matrix of H will have the form

$$H = \hat{B}(\infty)^{-1} \begin{bmatrix} 1 & \cdots & 0 & & & & & \\ & & & \cdots & & & & \\ & & & & 1 & \cdots & 0 & \\ & & & & & & & 1 \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix}. \tag{8.5}$$

Here, we see that we will need the inverse of $\hat{B}(\infty)$. Finally, if we let $G(\lambda)$ denote the matrix whose columns are the images under G in $\mathbb{R}^k[\lambda]$ of the basis elements for $X^R/\lambda^{-1}N^R$ displayed above, then we can compute a similar matrix for F by the formula

$$F(\lambda) = \lambda G(\lambda) - R(\lambda)H, \tag{8.6}$$

which follows from the definitions of F, G, and H. This is easily transformed into a matrix expression for F because of the simple basis we chose for X_R.

Example 8.1. Let $R(s)$ be given by

$$R(s) = \begin{pmatrix} s^2 & s^2+1 & 0 \\ 1 & s+2 & 3 \end{pmatrix}. \tag{8.7}$$

The leading row coefficient matrix of this is

$$B(\infty) = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \tag{8.8}$$

which has full row rank, so $R(s)$ is already row proper. The row degrees are 2 and 1, so a polynomial basis matrix for X_R is given by

$$\begin{pmatrix} \lambda & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{8.9}$$

We get $G = [I\ \ 0] \in \mathbb{R}^{3\times 4}$. We now have to choose $\tilde{B}$ to complete $B(\infty)$ to an invertible matrix; we can take $\tilde{B} = [0\ \ 0\ \ 1]$, which gives

$$\hat{B}(\infty) = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{8.10}$$

so that

$$\hat{B}(\infty)^{-1} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{8.11}$$

Therefore,

$$H = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{8.12}$$

Finally,

$$F(\lambda)=\begin{pmatrix}\lambda^2 & \lambda & 0 & 0\\ 0 & 0 & \lambda & 0\end{pmatrix}-\begin{pmatrix}\lambda^2 & \lambda^2+1 & 0\\ 1 & \lambda+2 & 3\end{pmatrix}\begin{pmatrix}1 & 0 & -1 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}$$

$$=\begin{pmatrix}0 & \lambda & -1 & 0\\ -1 & 0 & -1 & -3\end{pmatrix}. \tag{8.13}$$

The matrix of F is, therefore,

$$F=\begin{pmatrix}0 & 1 & 0 & 0\\ 0 & 0 & -1 & 0\\ -1 & 0 & -1 & -3\end{pmatrix}. \tag{8.14}$$

This concludes the example.

Now, suppose that a division of the external variables into inputs and outputs has been given, and that we want to obtain a representation in descriptor form. We start from the autoregressive equations, which appear in partitioned form:

$$[R_1(\sigma)\quad R_2(\sigma)]\begin{pmatrix}y\\ u\end{pmatrix}=0. \tag{8.15}$$

Taking $R(s)$ to row proper form as before, we get a corresponding partitioning of the right bicausal matrix $B(s)$:

$$[R_1(s)\quad R_2(s)]=\Delta(s)[B_1(s)\quad B_2(s)]. \tag{8.16}$$

By renumbering the inputs if necessary, we may assume that

$$B_2(\infty)=[B_2^1(\infty)\quad B_2^2(\infty)] \tag{8.17}$$

where $B_2^1(\infty)$ has full column rank, and the columns of $B_2^2(\infty)$ depend linearly on those of $[B_1(\infty)\ B_2^1(\infty)]$. Let $B_2^2(\infty)$ have m_2 columns; note that $m_2 \leqq q-k$. It is easily verified that a matrix $\tilde{B}$ which completes $B(\infty)$ to an invertible matrix may be found whose last m_2 rows are in the form $[0\ I]$. By the construction, a basis matrix for $\ker[B_1(\infty)\ B_2^1(\infty)]$ must be of the form $[N\ 0]^\top$. Taking these facts together, we conclude that $\hat{B}(\infty)^{-1}$ is of the form

$$\hat{B}(\infty)^{-1}=\begin{pmatrix}* & * & *\\ * & 0 & *\\ 0 & 0 & I\end{pmatrix} \tag{8.18}$$

where the partitioning is $(p+m_1+m_2)\times(k+(q-k-m_2)+m_2)$ (p is the number of y-variables, m_1 is the number of columns of $B_2^1(\infty)$). We therefore obtain equations of the following form:

$$\sigma z_0 = A_0 z_0 + B_1 z_1 + B_2 z_2 \tag{8.19}$$

$$y = H_{00} z_0 + H_{01} z_1 + H_{02} z_2 \tag{8.20}$$

$$u_1 = H_{10} z_0 + H_{12} z_2 \tag{8.21}$$

$$u_2 = z_2. \tag{8.22}$$

This can obviously be rewritten as

$$\sigma\begin{pmatrix}I & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}z_0\\ z_1\end{pmatrix}=\begin{pmatrix}A_0 & B_1\\ H_{10} & 0\end{pmatrix}\begin{pmatrix}z_0\\ z_1\end{pmatrix}+\begin{pmatrix}0 & B_2\\ -I & H_{12}\end{pmatrix}\begin{pmatrix}u_1\\ u_2\end{pmatrix} \tag{8.23}$$

$$y=[H_{00}\quad H_{01}]\begin{pmatrix}z_0\\ z_1\end{pmatrix}+[0\quad H_{02}]\begin{pmatrix}u_1\\ u_2\end{pmatrix}. \tag{8.24}$$

We now have a representation in descriptor form; as can be verified by checking the dimensions (using Remark 7.7), it is in fact a minimal representation. The fact that a zero block appears in the bottom right corner of the "A-matrix" means that the system "has no nondynamic variables" ([17]). It will be shown in [9] that the absence of nondynamic variables is a necessary condition for minimality of descriptor representations under external equivalence.

Example 8.2. Take

$$R(s)=\begin{pmatrix} s+1 & 0 & s^2 & 2 \\ s+2 & 2s & 1 & s-1 \end{pmatrix} \tag{8.25}$$

and let the first two external variables be outputs, and the other two inputs. The leading row coefficient matrix

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{pmatrix}$$

has full row rank, so that the given matrix $R(s)$ is already row reduced; also, $m_2=1$ and the inputs need not be renumbered. We see that the sum of the row indices of $R(s)$ is 3 and that the row and the column defects of $B_1(\infty)$ (formed by the first two columns of the matrix above) are both equal to 1; so, a descriptor representation (E, A, B, C, D) will be minimal if and only if the matrix E has size 4×4 and rank 3.

We can take

$$\tilde{B}=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{8.26}$$

which leads to

$$\hat{B}(\infty)^{-1}=\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{8.27}$$

Consequently, we get

$$H=\begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{8.28}$$

The matrix of F is computed from

$$\begin{aligned} F(\lambda)=&\begin{pmatrix} \lambda^2 & \lambda & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 \end{pmatrix}-\begin{pmatrix} \lambda+1 & 0 & \lambda^2 & 2 \\ \lambda+2 & 2\lambda & 1 & \lambda-1 \end{pmatrix} \\ &\times\begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \\ =&\begin{pmatrix} 0 & \lambda & 0 & -\lambda-1 & -2 \\ -1 & 0 & 0 & 2 & 1 \end{pmatrix}. \end{aligned} \tag{8.29}$$

This gives

$$F = \begin{pmatrix} 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & -2 \\ -1 & 0 & 0 & 2 & 1 \end{pmatrix}. \tag{8.30}$$

Of course, $G = [I_3 \ 0]$. Reorganizing the pencil equations as described above, we obtain

$$\sigma \begin{pmatrix} I_3 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & -2 \\ 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \tag{8.31}$$

$$y = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{8.32}$$

9. Conclusions. In this paper, we have shown a procedure which leads from a representation in autoregressive form (and in particular, from a left polynomial factorization) to a minimal descriptor representation. This procedure does not require the separation of finite and infinite frequencies. In fact, the transfer matrix is never computed, and the heaviest computational load in the algorithm consists of the inversion of a single constant matrix. The basic tool that we used is the pencil representation, which appears as a natural form that can be derived from autoregressive equations by a very simple formula. This formula also provides the link between the realization theory of Willems and that of Fuhrmann. The direct connection between autoregressive representations and descriptor representations which has now been established enables us to study more closely the relations between the two representations.

REFERENCES

[1] J. D. APLEVICH, *Time-domain input-output representations of linear systems*, Automatica, 17 (1981), pp. 509–522.

[2] ———, *Minimal representations of implicit linear systems*, Automatica, 21 (1985), pp. 259–269.

[3] H. BLOMBERG AND R. YLINEN, *Algebraic Theory for Multivariable Linear Systems*, Academic Press, London, 1983.

[4] G. CONTE AND A. M. PERDON, *Generalized state-space realizations for non-proper rational transfer functions*, Systems Control Lett., 1 (1982), pp. 270–276.

[5] P. A. FUHRMANN, *Algebraic system theory: an analyst's point of view*, J. Franklin Inst., 301 (1976), pp. 521–540.

[6] ———, *Linear Systems and Operators in Hilbert Space*, McGraw-Hill, New York, 1981.

[7] J. GRIMM, *Application de la théorie des systèmes implicites a l'inversion des systèmes*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Proc. 6th Int. Conf., Nice, June 1984; part 2, Lecture Notes Control Information Sciences 63, Springer-Verlag, Berlin, New York, 1984, pp. 142–156.

[8] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[9] M. KUIJPER AND J. M. SCHUMACHER, *Minimality of descriptor representations under external equivalence*, Report BS-R9002, CWI, Amsterdam, 1990.

[10] D. G. LUENBERGER, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, AC-22, (1977), pp. 312–321.

[11] D. G. LUENBERGER AND A. ARBEL, *Singular dynamic Leontief models*, Econometrica, 45 (1978), pp. 473–481.

[12] H. H. ROSENBROCK, *Structural properties of linear dynamical systems*, Internat. J. Control, 20 (1974), pp. 191–202.

[13] H. H. ROSENBROCK, *Non-minimal* LCR *multiports*, Internat. J. Control, 20 (1974), pp. 1–16.

[14] D. SALAMON, *Infinite dimensional systems with unbounded control and observation: a functional analytic approach*, Trans. AMS, 300 (1987), pp. 383–431.

[15] J. M. SCHUMACHER, *Transformations of linear systems under external equivalence*, Linear Algebra Appl., 102 (1988), pp. 1-34.

[16] P. VAN DOOREN, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 111-129.

[17] G. C. VERGHESE, B. LÉVY, AND T. KAILATH, *A generalized state space for singular systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 811-831.

[18] J. C. WILLEMS, *Input-output and state-space representations of finite-dimensional linear time-invariant systems*, Linear Algebra Appl., 50 (1983), pp. 581-608.

[19] ———, *From time series to linear system. Part* I: *Finite dimensional linear time invariant systems*, Automatica, 22 (1986), pp. 561-580.

[20] ———, *From time series to linear system. Part* II: *Exact modelling*, Automatica, 22 (1986), pp. 675-694.

[21] ———, *From time series to linear system. Part* III: *Approximate modelling*, Automatica, 23 (1987), pp. 87-115.

[22] H. K. WIMMER, *The structure of nonsingular polynomial matrices*, Math. Systems Theory, 14 (1981), pp. 367-379.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 1190-1208, September 1990

009

THE QUADRATIC MATRIX INEQUALITY IN SINGULAR H_∞ CONTROL WITH STATE FEEDBACK*

A. A. STOORVOGEL† AND H. L. TRENTELMAN†

Abstract. In this paper the standard H_∞ control problem using state feedback is considered. Given a linear, time-invariant, finite-dimensional system, this problem consists of finding a static state feedback such that the resulting closed-loop transfer matrix has H_∞ norm smaller than some a priori given upper bound. In addition it is required that the closed-loop system is internally stable. Conditions for the existence of a suitable state feedback are formulated in terms of a quadratic matrix inequality, reminiscent of the dissipation inequality of singular linear quadratic optimal control. Where the direct feedthrough matrix of the control input is injective, the results presented here specialize to known results in terms of solvability of a certain indefinite algebraic Riccati equation.

Key words. H_∞ control, state feedback, quadratic matrix inequality, strong controllability, almost disturbance decoupling

AMS(MOS) subject classifications. 93C05, 93C35, 93C45, 93C60, 93B27, 49B99

1. Introduction.

In a series of recent papers [1], [2], [5], [8], [10], [15], [18], [23] the by now well-known H_∞ optimal control problem was studied in a perspective of classical linear quadratic optimal control theory. In these papers it is shown that the existence of feedback controllers that result in a closed-loop transfer matrix with H_∞ norm less than a given upper bound is equivalent to the existence of solutions of certain algebraic Riccati equations. Typically, these algebraic Riccati equations are of the type we encounter in the context of linear quadratic differential games.

The first contributions to this new approach in H_∞ optimal control theory were reported in [8], [10], and [23]. These papers deal with the special case where the controllers to be designed are restricted to being state feedback control laws. In later contributions [2], [5], [18] these results were extended to the more general case of dynamic measurement feedback.

If we take a close look at the *type* of conditions for the existence of suitable controllers that are derived in the above references, we see there is a fundamental distinction between two cases. This distinction is tied up with the question of whether or not the *direct feedthrough matrix* of the control input is *injective.* In [10] and [23], *no* assumptions are imposed on the direct feedthrough matrix. The conditions for the existence of a suitable state feedback control law are formulated in terms of a *family* of algebraic Riccati equations, parameterized by a positive real parameter ε. It is shown that there exists an internally stabilizing state feedback control law such that the closed-loop transfer matrix has H_∞ norm less than an a priori given upper bound if and only if there exists a parameter value ε for which the corresponding Riccati equation has a certain solution. In our opinion, a more satisfactory type of condition is obtained in [2], [5], and [18]. In these papers it is assumed that the direct feedthrough matrix of the control input is *injective.* It is then shown that a suitable state feedback control law exists if and only if *one particular* algebraic Riccati equation has a solution with certain properties.

The purpose of the present paper is to reexamine the H_∞ problem with state feedback as studied in [2] and [18], *without making the assumption that the above-mentioned direct feedthrough matrix is injective.* Our aim is to find conditions for the

* Received by the editors March 6, 1989; accepted for publication (in revised form) November 7, 1989.
† Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands.

existence of suitable state feedback control laws that are of a different type from the one derived in [8], [10], and [23]. Instead our conditions will be of the type proposed in [2] and [18]. Stated differently: we will show how it is possible "to get rid of the parameter ε" in the conditions for the existence of suitable state feedback control laws. Rather than in terms of a particular algebraic Riccati equation, our conditions will be in terms of a certain "quadratic matrix inequality," reminiscent of the dissipation inequality appearing in singular linear quadratic optimal control [4], [13], [19]. It will turn out that the results from [2] and [18] on the special case that the direct feedthrough matrix is injective can be re-obtained from our results.

The outline of this paper is as follows. In § 2 we introduce the problem to be studied and give a statement of our main result. In § 3 we recall some important notions that will be used in this paper. In § 4 we give a description of a decomposition of the input space, the state space and the output space. This decomposition will be instrumental in the proof of our main result. Sections 5 and 6 are devoted to a proof of our main result. Finally, the paper closes with a brief discussion on our results in § 7.

2. Problem formulation and main results. We consider the finite-dimensional, linear, time-invariant system

$$\dot{x} = Ax + Bu + Ew, \qquad z = Cx + Du, \tag{2.1}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the control input, $w \in \mathbb{R}^l$ is an unknown disturbance, and $z \in \mathbb{R}^p$ is the output to be controlled. A, B, C, D, and E are real matrices of appropriate dimensions. In this paper we are primarily interested in state feedback. If F is a real $m \times n$ matrix, then the closed-loop transfer matrix resulting from the state feedback control law $u = Fx$ is equal to

$$G_F(s) = (C + DF)(Is - A - BF)^{-1}E.$$

The influence of the disturbance w on the output z is measured by the H_∞ norm of this transfer matrix:

$$\|G_F\|_\infty := \sup_{\omega \in \mathbb{R}} \rho[G_F(i\omega)].$$

Here, $\rho[M]$ denotes the largest singular value of the complex matrix M. The problem that we will study in this paper is the following: given a positive real number γ, find $F \in \mathbb{R}^{m \times n}$ such that

$$\|G_F\|_\infty < \gamma \quad \text{and} \quad \sigma(A + BF) \subset \mathbb{C}^-.$$

Here, $\sigma(M)$ denotes the set of eigenvalues of the matrix M and

$$\mathbb{C}^- := \{s \in \mathbb{C} \mid \operatorname{Re} s < 0\}.$$

A central role in our study of the above problem is played by what we will call the *quadratic matrix inequality*. For any real number $\gamma > 0$ and matrix $P \in \mathbb{R}^{n \times n}$ we define a matrix $F_\gamma(P) \in \mathbb{R}^{(n+m)\times(n+m)}$ by

$$F_\gamma(P) := \begin{pmatrix} PA + A^TP + \gamma^{-2}PEE^TP + C^TC & PB + C^TD \\ B^TP + D^TC & D^TD \end{pmatrix}. \tag{2.2}$$

Clearly, if P is symmetric, then $F_\gamma(P)$ is symmetric as well. If $F_\gamma(P) \geqq 0$, then we will say that *P is a solution to the quadratic matrix inequality at γ.*

In addition to (2.2), for any $\gamma > 0$ and $P \in \mathbb{R}^{n \times n}$ we define a $n \times (n+m)$ polynomial matrix $L_\gamma(P, s)$ by

$$L_\gamma(P, s) := (sI_n - A - \gamma^{-2}EE^TP \quad -B). \tag{2.3}$$

We note that $L_\gamma(P, s)$ is the controllability pencil associated with the system

$$\dot{x} = (A + \gamma^{-2} EE^T P)x + Bu.$$

The transfer matrix of the system Σ given by the equations

$$\dot{x} = Ax + Bu, \qquad y = Cx + Du \tag{2.4}$$

is equal to the real rational $p \times m$ matrix $G(s) = C(Is - A)^{-1}B + D$. The *normal rank* of a real rational matrix is defined as its rank as a matrix with entries in the field of real rational functions. The normal rank of the transfer matrix G is denoted by normrank G.

In the formulation of our main result we need the concept of *invariant zero* of the system $\Sigma = (A, B, C, D)$. For this definition we refer to § 3 (see also [11]). Finally, let $\mathbb{C}^0 := \{s \in \mathbb{C} \mid \text{Re } s = 0\}$ and let $\mathbb{C}^+ := \{s \in \mathbb{C} \mid \text{Re } s > 0\}$. The following is the main result of this paper.

THEOREM 2.1. *Consider the system* (2.1). *Assume that* (A, B, C, D) *has no invariant zeros in* $\mathbb{C}^0$. *Let* $\gamma > 0$. *Then the following two statements are equivalent*:

(i) *There exists* $F \in \mathbb{R}^{m \times n}$ *such that* $\|G_F\|_\infty < \gamma$ *and* $\sigma(A + BF) \subset \mathbb{C}^-$.

(ii) *There exists a real symmetric solution* $P \geqq 0$ *to the quadratic matrix inequality at* γ *such that*

$$\text{rank } F_\gamma(P) = \text{normrank } G \tag{2.5}$$

and

$$\text{rank} \begin{pmatrix} L_\gamma(P, s) \\ F_\gamma(P) \end{pmatrix} = n + \text{normrank } G \text{ for all } s \in \mathbb{C}^0 \cup \mathbb{C}^+. \tag{2.6}$$

In other words, the existence of a suitable state feedback control law is equivalent to the existence of a particular positive semidefinite solution of the quadratic matrix inequality at γ. This solution should be such that two rank conditions are satisfied.

Before embarking on a proof of this theorem we would like to point out how the results from [2] and [18] for the special case that D is injective can be obtained from our theorem as a special case. First note that in this case we have

$$\text{normrank } G = m.$$

Define

$$R_\gamma(P) := PA + A^T P + \gamma^{-2} PEE^T P + CC^T - (PB + C^T D)(D^T D)^{-1}(B^T P + D^T C).$$

Furthermore, define a real $(n+m) \times (n+m)$ matrix by

$$S(P) := \begin{pmatrix} I_n & -(PB + C^T D)(D^T D)^{-1} \\ 0 & I_m \end{pmatrix}.$$

Then clearly we have

$$S(P) F_\gamma(P) S(P)^T = \begin{pmatrix} R_\gamma(P) & 0 \\ 0 & D^T D \end{pmatrix}.$$

From this we can see that the pair of conditions $F_\gamma(P) \geqq 0$, rank $F_\gamma(P) = m$ is equivalent to the single condition $R_\gamma(P) = 0$. We now analyze the second rank condition appearing in our theorem. It is easily verified that for all $s \in \mathbb{C}$ we have

$$\begin{pmatrix} I_n & 0 & B(D^TD)^{-1} \\ 0 & I_n & -(PB + C^TD)(D^TD)^{-1} \\ 0 & 0 & I_m \end{pmatrix} \begin{pmatrix} L_\gamma(P, s) \\ F_\gamma(P) \end{pmatrix}$$
$$= \begin{pmatrix} sI - A - \gamma^{-2}EE^TP + B(D^TD)^{-1}(B^TP + D^TC) & 0 \\ R_\gamma(P) & 0 \\ B^TP + D^TC & D^TD \end{pmatrix}.$$

Consequently, if $R_\gamma(P) = 0$ then the condition

$$\operatorname{rank} \begin{pmatrix} L_\gamma(P, s) \\ F_\gamma(P) \end{pmatrix} = n + m \quad \text{for all } s \in \mathbb{C}^0 \cup \mathbb{C}^+$$

is equivalent to

$$\operatorname{rank} (sI - A - \gamma^{-2}EE^TP + B(D^TD)^{-1}(B^TP + D^TC)) = n \quad \text{for all } s \in \mathbb{C}^0 \cup \mathbb{C}^+$$

or, equivalently,

$$\sigma(A + \gamma^{-2}EE^TP - B(D^TD)^{-1}(B^TP + D^TC)) \subset \mathbb{C}^-.$$

Thus, for the special case that the direct feedthrough matrix D is injective our main result specializes to Corollary 2.2.

COROLLARY 2.2. *Consider the system* (2.1) *with* D *injective. Assume that* (A, B, C, D) *has no invariant zeros in* $\mathbb{C}^0$. *Let* $\gamma > 0$. *Then the following two statements are equivalent*:

(i) *There exists* $F \in \mathbb{R}^{m \times n}$ *such that* $\|G_F\|_\infty < \gamma$ *and* $\sigma(A + BF) \subset \mathbb{C}^-$.

(ii) *There exists a real symmetric solution* $P \geqq 0$ *to the algebraic Riccati equation*

$$PA + A^TP + \gamma^{-2}PEE^TP + CC^T - (PB + C^TD)(D^TD)^{-1}(B^TP + D^TC) = 0$$

such that

$$\sigma(A + \gamma^{-2}EE^TP - B(D^TD)^{-1}(B^TP + D^TC)) \subset \mathbb{C}^-.$$

A similar result was obtained in [2] and [18] for the special case that $D^TC = 0$ and $D^TD = I_m$. Our result differs slightly from those in [2] and [18] in the sense that we only require P to be semidefinite instead of definite.

3. Preliminaries and notation. In this section we recall some important notions that will be used in the sequel. First, we recall some facts about polynomial matrices. Let $\mathbb{R}[s]$ denote the ring of polynomials with real coefficients. Let $\mathbb{R}^{n \times m}[s]$ be the set of all $n \times m$ matrices with coefficients in $\mathbb{R}[s]$. An element of $\mathbb{R}^{n \times m}[s]$ is called a polynomial matrix. A square polynomial matrix is called unimodular if it is invertible. Two polynomial matrices P and Q are called unimodularly equivalent if there exist unimodular matrices U and V such that $Q = UPV$. In this paper, if P and Q are unimodularly equivalent, then we denote $P \sim Q$. It is well known [3] that for any

$P \in \mathbb{R}^{n\times m}[s]$ there exists $\Psi \in \mathbb{R}^{n\times m}[s]$ of the form

$$\Psi = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & & \vdots \\ \vdots & & \ddots & \psi_r & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & \vdots & & \vdots \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}$$

with ψ_i monic polynomials with the property that ψ_i divides ψ_{i+1}, such that $P \sim \Psi$. The polynomial matrix Ψ is called the *Smith form* of P (see [3]). The polynomials ψ_i are called the invariant factors of P. Their product $\psi := \psi_1\psi_2 \cdots \psi_r$ is called the zero polynomial of P. The roots of ψ are called the zeros of P. The integer r is equal to the normal rank of P; i.e., $r = \text{normrank } P$. If s is a complex number then $P(s)$ is an element of $\mathbb{C}^{n\times m}$. Its rank is denoted by rank $P(s)$. It is easy to see that normrank $P = \text{rank } P(s)$ for all $s \in \mathbb{C}$ if and only if P is unimodularly equivalent to the constant $n \times m$ matrix

$$\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix},$$

where I_r is the $r \times r$ identity matrix.

Next, we recall some important facts on the structure of the linear system Σ given by the equations (2.4). As before, this system is denoted by (A, B, C, D) or simply by Σ. The system matrix of Σ is defined as the polynomial matrix

$$P_\Sigma(s) = \begin{pmatrix} Is - A & -B \\ C & D \end{pmatrix}.$$

The invariant factors of P_Σ are called the *transmission polynomials* of Σ. The zeros of P_Σ are called the *invariant zeros of* Σ. Clearly, $s \in \mathbb{C}$ is an invariant zero of Σ if and only if rank $P_\Sigma(s) < \text{normrank } P_\Sigma$. It is easy to see that if $F \in \mathbb{R}^{m\times n}$ and if we define $\Sigma_F := (A + BF, B, C + DF, D)$, then $P_\Sigma \sim P_{\Sigma_F}$. In particular this implies that the transmission polynomials of Σ and Σ_F coincide and a fortiori that the invariant zeros of Σ and Σ_F coincide. An important role in this paper is played by *the strongly controllable subspace of* Σ. Consider the following sequence of subspaces:

$$\begin{aligned} &\mathcal{T}_0(\Sigma) = 0, \\ &\mathcal{T}_{i+1}(\Sigma) = \{x \in \mathbb{R}^n \mid \exists w \in \mathcal{T}_i(\Sigma),\ u \in \mathbb{R}^m \text{ s.t. } Aw + Bu = x \text{ and } Cw + Du = 0\}. \end{aligned} \tag{3.1}$$

It is well known (see [7]) that $\mathcal{T}_i(\Sigma)$ $(i = 1, 2, \cdots)$ is a nondecreasing sequence that attains its limit in finitely many steps. The limiting subspace is denoted by $\mathcal{T}(\Sigma)$ and is called the *strongly controllable subspace of* Σ. $\mathcal{T}(\Sigma)$ is known to be the smallest subspace $\mathcal{V}$ of $\mathbb{R}^n$ with the property that there exists a linear mapping G from $\mathbb{R}^p$ to $\mathbb{R}^n$ such that $(A + GC)\mathcal{V} \subseteq \mathcal{V}$ and $\text{im}\,(B + GD) \subseteq \mathcal{V}$. From this it is easily seen that $\mathcal{T}(\Sigma)$ is $(C + DF, A + BF)$-invariant for every linear mapping $F: \mathbb{R}^n \to \mathbb{R}^m$ (a subspace $\mathcal{V}$ is called (C, A)-invariant if it satisfies $A(\mathcal{V} \cap \ker C) \subseteq \mathcal{V}$; see also [12]). The system Σ is called *strongly controllable if* $\mathcal{T}(\Sigma) = \mathbb{R}^n$. If Σ is strongly controllable, then (A, B) is controllable. It is known that Σ is strongly controllable if and only if rank $P_\Sigma(s) = n + \text{rank}\,(C\ D)$ for every $s \in \mathbb{C}$ (see [6], [14]). Hence, by the above we find that if $(C\ D)$ is surjective, then Σ is strongly controllable if and only if P_Σ is unimodularly equivalent to the constant matrix $(I_{n+p}\ 0)$, where I_{n+p} denotes the $(n+p)\times(n+p)$ identity matrix.

We conclude this section by introducing some notation. We will denote $\mathbb{R}^+ := [0, \infty)$. $\mathscr{L}_2(\mathbb{R}^+)$ denotes the space of real-valued measurable functions from $\mathbb{R}^+$ to $\mathbb{R}$ such that $\int_{\mathbb{R}^+} \|x\|^2\, dt < \infty$. For a given positive integer r we denote by $\mathscr{L}_2^r(\mathbb{R}^+)$ the space of r-vectors with components in $\mathscr{L}_2(\mathbb{R}^+)$. The notation $\|\ \|$ is used for the Euclidean norm on $\mathbb{R}^r$; $\|\ \|_2$ denotes the usual norm on $\mathscr{L}_2^r(\mathbb{R}^+)$; i.e., $\|x\|_2 := (\int_{\mathbb{R}^+} \|x\|^2\, dt)^{1/2}$.

4. A preliminary feedback transformation. In this section we show that by applying a suitable state feedback transformation $u = F_0 x + v$ to the system $\Sigma = (A, B, C, D)$, it is transformed into a system $\Sigma_{F_0} := (A + BF_0, B, C + DF_0, D)$ with a very particular structure. We will display this structure by writing down the matrices of the mappings $A + BF_0$, B, $C + DF_0$, and D with respect to suitable bases in the input space $\mathbb{R}^m$, the state space $\mathbb{R}^n$, and the output space $\mathbb{R}^p$.

First choose a basis of $\mathbb{R}^m$ as follows. Let $q_1, \cdots, q_l, q_{l+1}, \cdots, q_m$ be a basis such that $q_{l+1}, \cdots, q_m$ is a basis of $\ker D$ $(0 \leqq 1 \leqq m)$. In other words, decompose $\mathbb{R}^m = \mathscr{U}_1 \oplus \mathscr{U}_2$, with $\mathscr{U}_2 = \ker D$ and $\mathscr{U}_1$ arbitrary. Next, choose a basis of $\mathbb{R}^p$ as follows. Let $z_1, \cdots, z_r, z_{r+1}, \cdots, z_p$ be an *orthonormal* basis such that $z_1, \cdots, z_r$ is an orthonormal basis of $\operatorname{im} D$ and $z_{r+1}, \cdots, z_p$ is an orthonormal basis of $(\operatorname{im} D)^\perp$ $(0 \leqq r \leqq p)$. In other words, write $\mathbb{R}^p = \mathscr{Z}_1 \oplus \mathscr{Z}_2$ with $\mathscr{Z}_1 = \operatorname{im} D$ and $\mathscr{Z}_2 = (\operatorname{im} D)^\perp$. If $\binom{z_1}{z_2}$ is the coordinate vector of a given $z \in \mathbb{R}^p$, then because of orthonormality we have $\|z\| = \|\binom{z_1}{z_2}\|$ (here $\|\ \|$ denotes the Euclidean norm). With respect to these decompositions the mapping D has the form

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix},$$

with D_1 invertible. Moreover, B and C can be partitioned as

$$B = (B_1 \quad B_2), \qquad C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}.$$

It is easy to see that $\operatorname{im} B_2 = B \ker D$ and $\ker C_2 = C^{-1} \operatorname{im} D := \{x \mid Cx \in \operatorname{im} D\}$.

Next, define a linear mapping $F_0 : \mathbb{R}^n \to \mathbb{R}^m$ by

$$F_0 := \begin{pmatrix} -D_1^{-1} C_1 \\ 0 \end{pmatrix}. \tag{4.1}$$

Then we have

$$C + DF_0 = \begin{pmatrix} 0 \\ C_2 \end{pmatrix}.$$

We now choose a basis of $\mathbb{R}^n$. Let $x_1, \cdots, x_s, x_{s+1}, \cdots, x_t, x_{t+1}, \cdots, x_n$ $(0 \leqq s \leqq t \leqq n)$ be a basis such that $x_{s+1}, \cdots, x_t$ is a basis of $\mathscr{T}(\Sigma) \cap C^{-1} \operatorname{im} D$ and $x_{s+1}, \cdots, x_n$ is a basis of $\mathscr{T}(\Sigma)$. In other words, write $\mathbb{R}^n = \mathscr{X}_1 \oplus \mathscr{X}_2 \oplus \mathscr{X}_3$ with $\mathscr{X}_2 = \mathscr{T}(\Sigma) \cap C^{-1} \operatorname{im} D$, $\mathscr{X}_2 \oplus \mathscr{X}_3 = \mathscr{T}(\Sigma)$ and $\mathscr{X}_1$ arbitrary. It turns out that with respect to the bases introduced above, $A + BF_0$, B and $C + DF_0$ have a particular form. This is a consequence of the following lemma.

LEMMA 4.1. *Let F_0 be given by* (4.1). *Then we have*:

(i) $(A + BF_0)(\mathscr{T}(\Sigma) \cap C^{-1} \operatorname{im} D) \subseteq \mathscr{T}(\Sigma)$,

(ii) $\operatorname{im} B_2 \subseteq \mathscr{T}(\Sigma)$,

(iii) $\mathscr{T}(\Sigma) \cap C^{-1} \operatorname{im} D \subseteq \ker C_2$.

Proof. (i) $\mathscr{T}(\Sigma)$ is $(C + DF_0, A + BF_0)$-invariant. This implies that

$$(A + BF_0)(\mathscr{T}(\Sigma) \cap \ker (C + DF_0)) \subseteq \mathscr{T}(\Sigma).$$

Since $\ker (C+DF_0) = \ker C_2 = C^{-1} \operatorname{im} D$, the result follows.

(ii) Let $\mathcal{T}_i(\Sigma)$ be the sequence defined by (3.1). Then $\mathcal{T}_1(\Sigma) = B \ker D = \operatorname{im} B_2$. Since $\mathcal{T}_i(\Sigma)$ is nondecreasing this proves our claim.

(iii) This follows immediately from the fact that $C^{-1} \operatorname{im} D = \ker C_2$.

By applying this lemma we find that the matrices of $A+BF_0$, B, $C+DF_0$, and D with respect to the given bases have the following form:

$$A+BF_0 = \begin{pmatrix} A_{11} & 0 & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \\ B_{31} & B_{32} \end{pmatrix},$$

$$C+DF_0 = \begin{pmatrix} 0 & 0 & 0 \\ C_{21} & 0 & C_{23} \end{pmatrix}, \qquad D = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix}. \tag{4.2}$$

If we apply the feedback transformation $u = F_0 x + v$ to the system $\Sigma = (A, B, C, D)$, then the resulting system Σ_{F_0} is given by

$$\dot{x} = (A+BF_0)x + Bv, \qquad z = (C+DF_0)x + Dv. \tag{4.3}$$

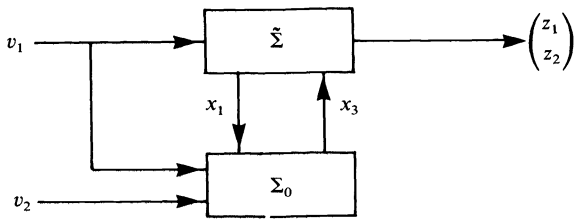
With respect to the given decomposition, let $\binom{v_1}{v_2}$ be the coordinate vector of a given $v \in \mathbb{R}^m$. Likewise, we use the notation $(x_1^T, x_2^T, x_3^T)^T$ and $\binom{z_1}{z_2}$. Then the equations of the system Σ_{F_0} can be arranged in such a way that they have the following form:

$$\dot{x} = A_{11}x_1 + \begin{pmatrix} B_{11} & A_{13} \end{pmatrix} \begin{pmatrix} v_1 \\ x_3 \end{pmatrix}, \tag{4.4}$$

$$\begin{pmatrix} \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} B_{22} \\ B_{32} \end{pmatrix} v_2 + \begin{pmatrix} B_{21} & A_{21} \\ B_{31} & A_{31} \end{pmatrix} \begin{pmatrix} v_1 \\ x_1 \end{pmatrix}, \tag{4.5}$$

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 0 \\ C_{21} \end{pmatrix} x_1 + \begin{pmatrix} D_1 & 0 \\ 0 & C_{23} \end{pmatrix} \begin{pmatrix} v_1 \\ x_3 \end{pmatrix}. \tag{4.6}$$

As already suggested by the way that we have arranged these equations, the system Σ_{F_0} can be considered as the interconnection of two subsystems. This is depicted as follows:



Here,

$$\tilde{\Sigma} := \left(A_{11}, \begin{pmatrix} B_{11} & A_{13} \end{pmatrix}, \begin{pmatrix} 0 \\ C_{21} \end{pmatrix}, \begin{pmatrix} D_1 & 0 \\ 0 & C_{23} \end{pmatrix} \right) \tag{4.7}$$

is the system given by (4.4) and (4.6). It has input space $\mathcal{U}_1 \times \mathcal{X}_3$, state space $\mathcal{X}_1$, and output space $\mathbb{R}^p$. Σ_0 is the system given by (4.5). It has input space $\mathbb{R}^m \times \mathcal{X}_1$ and state space $\mathcal{X}_2 \oplus \mathcal{X}_3$. The interconnection is made via x_1 and x_3, as depicted above. Note that $\tilde{\Sigma}$ and Σ_{F_0} have the same output equation. However, in Σ_{F_0} the variable x_3 is

generated by Σ_0, whereas in $\tilde{\Sigma}$ it is considered as an input and is free. The systems Σ_0 and $\tilde{\Sigma}$ turn out to have a couple of nice structural properties, as shown in Lemma 4.2.

LEMMA 4.2. (i) C_{23} *is injective,*

(ii) *The system*

$$\Sigma_1 := \left(\begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{pmatrix}, \begin{pmatrix} B_{22} \\ B_{32} \end{pmatrix}, (0 \quad I), 0 \right) \tag{4.8}$$

with input space $\mathcal{U}_2$, *state space* $\mathcal{X}_2 \oplus \mathcal{X}_3$ $(= \mathcal{T}(\Sigma))$, *and output space* $\mathcal{X}_3$ *is strongly controllable.*

Proof. (i) Let $(x_1^T, x_2^T, x_3^T)^T$ be the coordinate vector of a given $x \in \mathbb{R}^n$. Assume that $C_{23}x_3 = 0$. Let $\tilde{x} \in \mathbb{R}^n$ be the vector with coordinates $(0^T, 0^T, x_3^T)^T$. Then $\tilde{x} \in \mathcal{X}_3$. In addition, $\tilde{x} \in \mathcal{T}(\Sigma) \cap \ker C_2 = \mathcal{X}_2$. Thus $\tilde{x} = 0$, so $x_3 = 0$.

(ii) Let $\mathcal{T}(\Sigma_1)$ be the strongly controllable subspace of the system Σ_1 given by (4.8). We will prove that $\mathcal{T}(\Sigma_1) = \mathcal{X}_2 \oplus \mathcal{X}_3$. First note that there exists $G = \binom{G_2}{G_3}$ such that

$$\left(\begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{pmatrix} + \begin{pmatrix} G_2 \\ G_3 \end{pmatrix} (0 \quad I) \right) \mathcal{T}(\Sigma_1) \subseteq \mathcal{T}(\Sigma_1).$$

Also note that

$$\operatorname{im} \begin{pmatrix} B_{22} \\ B_{32} \end{pmatrix} \subseteq \mathcal{T}(\Sigma_1).$$

Now assume that $\mathcal{T}(\Sigma_1) \subseteq \mathcal{X}_2 \oplus \mathcal{X}_3$ with *strict* inclusion. Define $\mathcal{V} \subseteq \mathbb{R}^n$ by

$$\mathcal{V} := \left\{ \begin{pmatrix} 0 \\ x_2 \\ x_3 \end{pmatrix} \middle| \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} \in \mathcal{T}(\Sigma_1) \right\}.$$

Clearly, $\mathcal{V} \subseteq \mathcal{T}(\Sigma)$ with strict inclusion. We claim that there exists a linear map $G_0 : \mathbb{R}^p \to \mathbb{R}^n$ such that

$$(A + G_0 C)\mathcal{V} \subseteq \mathcal{V} \tag{4.9}$$

$$\operatorname{im} (B + G_0 D) \subseteq \mathcal{V}. \tag{4.10}$$

Indeed, let C_{23}^+ be any left inverse of C_{23} and define

$$G_0 := \begin{pmatrix} B_{11} & -A_{13} \\ B_{21} & G_2 \\ B_{31} & G_3 \end{pmatrix} \begin{pmatrix} -D_1^{-1} & 0 \\ 0 & C_{23}^+ \end{pmatrix}.$$

It is then straightforward to verify (4.9) and (4.10). This, however, contradicts the fact that $\mathcal{T}(\Sigma)$ is the smallest subspace $\mathcal{V}$ for which (4.9) and (4.10) hold (see § 3). We conclude that $\mathcal{X}_2 \oplus \mathcal{X}_3 = \mathcal{T}(\Sigma_1)$. □

Our next result states that the zero structure of the original system $\Sigma = (A, B, C, D)$ is completely determined by the zero structure of the subsystem $\tilde{\Sigma}$ given by (4.7). A transmission polynomial of a system is called *nontrivial* if it is unequal to the constant polynomial 1.

LEMMA 4.3. *The nontrivial transmission polynomials of* Σ *and* $\tilde{\Sigma}$, *respectively, coincide.*

Proof. According to § 3 the transmission polynomials of Σ and Σ_{F_0} coincide. Thus, to prove the lemma it suffices to show that the system matrix P_0 of Σ_{F_0} is unimodularly

equivalent to a polynomial matrix of the form

$$\begin{pmatrix} P_{\tilde{\Sigma}}(s) & 0 & 0 \\ 0 & I & 0 \end{pmatrix},$$

where $P_{\tilde{\Sigma}}(s)$ is the system matrix of $\tilde{\Sigma}$. Since Σ_1 is strongly controllable and $(0\ \ I)$ is surjective, the Smith form of P_{Σ_1} is equal to $(I_1\ \ 0)$ (I_1 denotes the identity matrix with size equal to $\dim \mathscr{X}_2 + 2\dim \mathscr{X}_3$). In addition, clearly we have

$$P_{\Sigma_1} \sim \begin{pmatrix} sI - A_{22} & 0 & -B_{22} \\ -A_{32} & 0 & -B_{32} \\ 0 & I & 0 \end{pmatrix} \sim \begin{pmatrix} sI - A_{22} & -B_{22} & 0 \\ -A_{32} & -B_{32} & 0 \\ 0 & 0 & I \end{pmatrix},$$

so we conclude that

$$\begin{pmatrix} sI - A_{22} & -B_{22} \\ -A_{32} & -B_{32} \end{pmatrix}$$

is unimodularly equivalent to $(I_2\ \ 0)$. Here I_2 denotes the identity matrix of size $\dim \mathscr{X}_2 + \dim \mathscr{X}_3$. The proof is then completed by noting that

$$P_0 \sim \begin{pmatrix} sI - A_{11} & -B_{11} & -A_{13} & 0 & 0 \\ 0 & D_1 & 0 & 0 & 0 \\ C_{21} & 0 & C_{23} & 0 & 0 \\ -A_{21} & -B_{21} & -A_{23} & sI - A_{22} & -B_{22} \\ -A_{31} & -B_{31} & sI - A_{33} & -A_{32} & -B_{32} \end{pmatrix} \sim \begin{pmatrix} P_{\tilde{\Sigma}}(s) & 0 & 0 \\ 0 & I_2 & 0 \end{pmatrix}.$$

A consequence of the above lemma is that the invariant zeros of Σ and $\tilde{\Sigma}$, respectively, coincide.

Our next lemma states that the normal rank of the transfer matrix $G(s) = C(sI - A)^{-1}B + D$ of the system Σ is equal to the number $\operatorname{rank} D_1 + \dim \mathscr{X}_3$ or, equivalently, Lemma 4.4.

LEMMA 4.4. *We have*

$$\operatorname{normrank} G = \operatorname{rank} \begin{pmatrix} C_{23} & 0 \\ 0 & D_1 \end{pmatrix}.$$

Proof. Define $L(s) := sI - A$. Then we have

$$\operatorname{normrank} \begin{pmatrix} L & 0 \\ 0 & G \end{pmatrix} = n + \operatorname{normrank} G. \tag{4.11}$$

We also have

$$\begin{pmatrix} I & 0 \\ C(sI - A)^{-1} & I \end{pmatrix} \begin{pmatrix} L(s) & 0 \\ 0 & G(s) \end{pmatrix} \begin{pmatrix} I & -(sI - A)^{-1}B \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ F_0 & I \end{pmatrix}$$

$$= \begin{pmatrix} sI - (A + BF_0) & -B \\ C + DF_0 & D \end{pmatrix} = \begin{pmatrix} sI - A_{11} & 0 & -A_{13} & -B_{11} & 0 \\ -A_{21} & sI - A_{22} & -A_{23} & -B_{21} & -B_{22} \\ -A_{31} & -A_{32} & sI - A_{33} & -B_{31} & -B_{32} \\ 0 & 0 & 0 & D_1 & 0 \\ C_{21} & 0 & C_{23} & 0 & 0 \end{pmatrix}.$$

Since C_{23} and D_1 are injective, we can make the (1, 3), (1, 4), (2, 4), and (3, 4) blocks zero by unimodular transformations. Furthermore, we can make a basis transformation

on the output such that C_{23} has the form $\binom{I_r}{0}$ where $r = \operatorname{rank} C_{23}$. Thus, after suitable permutation of blocks, the normal rank of the latter matrix turns out to be equal to the normal rank of

$$\begin{pmatrix} sI-\tilde{A}_{11} & 0 & 0 & 0 & 0 \\ -A_{21} & sI-A_{22} & -A_{23} & -B_{22} & 0 \\ -A_{31} & -A_{32} & sI-A_{33} & -B_{32} & 0 \\ C_{211} & 0 & I_r & 0 & 0 \\ C_{212} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & D_1 \end{pmatrix}.$$

Here $\tilde{A}_{11}$ is a given matrix. Since, by Lemma 4.2, the matrix in the center has full row rank for all $s \in \mathbb{C}$ and since normrank $(sI - \tilde{A}_{11}) = \dim \mathscr{X}_1$, we find

$$\operatorname{normrank} \begin{pmatrix} L & 0 \\ 0 & G \end{pmatrix} = n + \operatorname{rank} \begin{pmatrix} C_{23} & 0 \\ 0 & D_1 \end{pmatrix}.$$

Combining this with (4.11), we obtain the desired result. □

To conclude this section we want to note that if D is injective, then the subspace $\mathscr{U}_2$ in the decomposition of $\mathbb{R}^m$ vanishes. Consequently, the partitioning of B reduces to a single block and the partitioning of D reduces to $\binom{D_1}{0}$ with D_1 invertible. It is left as an exercise to the reader to show that $\mathscr{T}(\Sigma) = 0$ if and only if $\ker D \subseteq \ker B$. Thus, if D is injective, then also $\mathscr{T}(\Sigma) = 0$. In that case the subspaces $\mathscr{X}_2$ and $\mathscr{X}_3$ appearing in the decomposition of $\mathscr{X}$ both vanish and the partitioning of $A + BF_0$ reduces to a single block.

5. Solvability of the quadratic matrix inequality. In this section we will establish a proof of the implication (i)⇒(ii) in Theorem 2.1: assuming that a suitable state feedback control law exists, we show that the quadratic matrix inequality has a solution with the asserted properties.

Consider our control system (2.1). For given disturbance and control functions w and u we denote by $x_{w,u}$ and $z_{w,u}$ the corresponding state trajectory and output function, respectively, with $x(0) = 0$. We will first formulate a theorem that serves as a basis for the developments in the rest of this paper. The theorem is concerned with the special case that in the system (2.1) the direct feedthrough matrix D is injective. The result in Theorem 5.1 is a generalization of [2, Thm. 2] and of results in [18].

THEOREM 5.1. *Consider the system* (2.1) *and assume that D is injective. Assume that (A, B, C, D) has no invariant zeros in $\mathbb{C}^0$. Let $\gamma > 0$. Then the following statements are equivalent*:

(i) *(A, B) is stabilizable and, in addition, there exists $\delta > 0$ such that for all $w \in \mathscr{L}_2^l(\mathbb{R}^+)$ there exists $u \in \mathscr{L}_2^m(\mathbb{R}^+)$ for which $x_{w,u} \in \mathscr{L}_2^n(\mathbb{R}^+)$ and $\|z_{w,u}\|_2 \leqq (\gamma - \delta)\|w\|_2$.*

(ii) *There exists a real symmetric solution $P \geqq 0$ to the algebraic Riccati equation*

$$PA + A^T P + \gamma^{-2} PEE^T P + C^T C - (PB + C^T D)(D^T D)^{-1}(B^T P + D^T C) = 0 \tag{5.1}$$

such that

$$\sigma(A + \gamma^{-2} EE^T P - B(D^T D)^{-1}(B^T P + D^T C)) \subset \mathbb{C}^-. \tag{5.2}$$

Moreover, if the latter holds, then one possible choice for u is given by $u = Fx$, *with*

$$F = -(D^TD)^{-1}(B^TP + D^TC).$$

For this F we have $\|G_F\|_\infty < \gamma$ *and* $\sigma(A+BF) \subset \mathbb{C}^-$.

Proof. A proof of this theorem can be based on the proof of [18, Thm. 2.1c]. In the latter paper it is assumed that C is injective and that $C^TD = 0$, which implies that (A, B, C, D) has no zeros at all. The proof of Theorem 2.1c of [18] can, however, be modified to yield a proof of our result. In doing this the following important point might need clarification. Since, in our context (C, A) is not necessarily detectable, we must make a careful distinction between the H_∞ problem with stability (i.e., $x \in \mathscr{L}_2^n$ and $u \in \mathscr{L}_2^m$) and the H_∞ problem without stability (i.e., no restrictions on x and u). In the proof of Theorem 2.1 of [18] a version of the maximum principle is used that gives a *sufficient* condition for optimality in the case that (C, A) is detectable (for a finite-horizon version of this result see [9, Chap. 5.2]). However, if we drop the detectability assumption, this method can still be used for the H_∞ problem with stability. The remainder of the proof in [18] can be checked step by step and remains valid.

Since in our context (C, A) is not necessarily observable (in contrast with [2] and [18]) our theorem involves a semidefinite solution of (5.1) rather than a definite one. □

Now, again consider the system (2.1), this time without making any assumptions on the matrix D. Choose bases in the state space, the input space, and the output space as in § 4 and apply the feedback transformation $u = F_0x + v$, with F_0 given by (4.1). After this transformation we have

$$\dot{x} = (A + BF_0)x + Bv + Ew, \qquad z = (C + DF_0)x + Dv. \tag{5.3}$$

If we partition $E = (E_1^T, E_2^T, E_3^T)^T$, then in terms of our decomposition (5.3) can be written as follows:

$$\dot{x}_1 = A_{11}x_1 + (B_{11} \quad A_{13})\begin{pmatrix} v_1 \\ x_3 \end{pmatrix} + E_1w, \tag{5.4}$$

$$\begin{pmatrix} \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{pmatrix}\begin{pmatrix} x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} B_{22} \\ B_{32} \end{pmatrix} v_2 + \begin{pmatrix} B_{21} & A_{21} \\ B_{31} & A_{31} \end{pmatrix}\begin{pmatrix} v_1 \\ x_1 \end{pmatrix} + \begin{pmatrix} E_2 \\ E_3 \end{pmatrix} w, \tag{5.5}$$

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 0 \\ C_{21} \end{pmatrix} x_1 + \begin{pmatrix} D_1 & 0 \\ 0 & C_{23} \end{pmatrix}\begin{pmatrix} v_1 \\ x_3 \end{pmatrix}. \tag{5.6}$$

For given disturbance and control functions w and v, let $x_{w,v}$ and $z_{w,v}$ denote the state trajectory and output, respectively, of (5.3), with $x(0) = 0$. The idea that we want to pursue is the following. If there exists a feedback law $u = Fx$ for (2.1) such that $\|G_F\|_\infty < \gamma$ and $\sigma(A+BF) \subset \mathbb{C}^-$, then the feedback law $v = (F - F_0)x$ in (5.3) yields a closed-loop transfer matrix from w to z with H_∞ norm smaller than γ. In other words,

$$\beta := \sup_{w \in \mathscr{L}_2(\mathbb{R}^+)} \frac{\|z_{w,v}\|_2}{\|w\|_2} < \gamma. \tag{5.7}$$

Also, $x_{w,v} \in \mathscr{L}_2^n(\mathbb{R}^+)$. Let $\delta := \gamma - \beta$. Then, for a given w, define v_1 as the first component of $v = (F - F_0)x_{w,v}$ and take x_3 as the third component of $x_{w,v}$. Interpret $\binom{v_1}{x_3}$ as an input for the subsystem $\tilde{\Sigma}$ defined by (5.4) and (5.6). It then follows from (5.7) that

$$\left\| \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \right\|_2 \leqq (\gamma - \delta)\|w\|_2.$$

Moreover, the "input" $\binom{v_1}{x_3}$ and the "state trajectory" x_1 are in $\mathscr{L}_2$. The crucial observation is now that the direct feedthrough matrix of $\tilde{\Sigma}$ is *injective* (see Lemma 4.2). Thus we can apply Theorem 5.1 to establish the existence of a solution to the algebraic Riccati equation associated with the system $\tilde{\Sigma}$. Before doing this, however, we should make sure that $(A_{11}, (B_{11}, A_{13}))$ is stabilizable and that $\tilde{\Sigma}$ given by (4.7) has no invariant zeros in $\mathbb{C}^0$. It is easily seen that if (A, B) is stabilizable, then also $(A_{11}, (B_{11}, A_{13}))$ is stabilizable. Furthermore, if $\Sigma = (A, B, C, D)$ has no invariant zeros in $\mathbb{C}^0$, then the same holds for $\tilde{\Sigma}$ (see Lemma 4.3). Consequently, we have the following corollary.

COROLLARY 5.2. *Consider the system* (2.1). *Assume that* (A, B, C, D) *has no invariant zeros in* $\mathbb{C}^0$. *Let* $\gamma > 0$ *and assume there exists* $F \in \mathbb{R}^{m\times n}$ *such that* $\|G_F\|_\infty < \gamma$ *and* $\sigma(A+BF) \subset \mathbb{C}^-$. *Then there exists a real symmetric solution* $P_{11} \geqq 0$ *to the algebraic Riccati equation*

$$\begin{aligned} &P_{11}A_{11} + A_{11}^T P_{11} + C_{21}^T C_{21} + \gamma^{-2} P_{11}E_1E_1^T P_{11} - P_{11}B_{11}(D_1^T D_1)^{-1}B_{11}^T P_{11} \\ &\quad -(A_{13}^T P_{11} + C_{23}^T C_{21})^T (C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}) = 0 \end{aligned} \tag{5.8}$$

such that

$$\begin{aligned} &\sigma(A_{11} + \gamma^{-2}E_1E_1^T P_{11} - B_{11}(D_1^T D_1)^{-1}B_{11}^T P_{11} \\ &\quad - A_{13}(C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21})) \subset \mathbb{C}^-. \end{aligned} \tag{5.9}$$

Our next step is to establish a connection between the algebraic Riccati equation (5.8) and the quadratic matrix inequality.

It turns out that there is a one-to-one correspondence between the set of solutions to (5.8) and the set of solutions to the quadratic matrix inequality at γ that satisfy the rank condition (2.5). To prove this, we need the following lemma.

LEMMA 5.3. *Assume* $P \in \mathbb{R}^{n\times n}$ *is a solution to the quadratic matrix inequality at* γ. *Then* $\mathscr{T}(\Sigma) \subseteq \ker P$.

Proof. Let F_0 be given by (4.1). Let $\mathscr{R}$ be the smallest $(C+DF_0, A+BF_0)$-invariant subspace containing $B \ker D$. We claim that $\mathscr{R} = \mathscr{T}(\Sigma)$. We know that $\mathscr{T}(\Sigma)$ is $(C+DF, A+BF)$-invariant for all F and hence also for $F = F_0$. Second, by Lemma 4.1(ii) we have $\mathscr{T}(\Sigma) \supseteq \ker D$. Therefore, we have $\mathscr{R} \subseteq \mathscr{T}(\Sigma)$. Conversely, we know that

$$\exists G_1 : \operatorname{im}(C+DF_0) \to \mathbb{R}^n \text{ s.t. } [(A+BF_0) + G_1(C+DF_0)]\mathscr{R} \subseteq \mathscr{R},$$

$$\exists G_2 : \operatorname{im} D \to \mathbb{R}^n \text{ s.t. } \operatorname{im}(B+G_2D) = B \ker D \subseteq \mathscr{R}.$$

Since $D^T(C+DF_0) = 0$ (this can be checked easily) we can find a linear mapping G such that $G|_{\operatorname{im}(C+DF_0)} = G_1$ and $G|_{\operatorname{im} D} = G_2$ and hence we have found a G such that $(A+GC)\mathscr{R} \subseteq \mathscr{R}$ and $\operatorname{im}(B+GD) \subseteq \mathscr{R}$. Thus we find $\mathscr{R} \supseteq \mathscr{T}(\Sigma)$ and hence $\mathscr{R} = \mathscr{T}(\Sigma)$.

Let $\gamma > 0$ and define

$$M_\gamma(P) := \begin{pmatrix} I & F_0^T \\ 0 & I \end{pmatrix} F_\gamma(P) \begin{pmatrix} I & 0 \\ F_0 & I \end{pmatrix}. \tag{5.10}$$

If $F_\gamma(P) \geqq 0$ then also

$$M_\gamma(P) = \begin{pmatrix} P(A+BF_0) + (A+BF_0)^T P + \gamma^{-2}PEE^TP + (C+DF_0)^T(C+DF_0) & PB \\ B^TP & D^TD \end{pmatrix} \geqq 0 \tag{5.11}$$

We claim that $B \ker D \subseteq \ker P$. Let $u \in \mathbb{R}^m$ be such that $Du = 0$. Then we find $\binom{0}{u}^T M_\gamma(P)\binom{0}{u} = 0$ and hence, since $M_\gamma(P) \geqq 0$, we find $M_\gamma(P)\binom{0}{u} = 0$. This implies $PBu = 0$. Next we have that $\ker P$ is $(C+DF_0, A+BF_0)$-invariant. Assume that $x \in$

$\ker P \cap \ker (C + DF_0)$. Then

$$x^T(P(A+BF_0)+(A+BF_0)^TP+\gamma^{-2}PEE^TP+(C+DF_0)^T(C+DF_0))x=0.$$

Hence, by applying x to one side only, we find $P(A+BF_0)x=0$ and therefore $(A+BF_0)x \in \ker P$. Since $\mathcal{T}(\Sigma)$ is the smallest space with these two properties, we must have $\mathcal{T}(\Sigma) \subseteq \ker P$. □

Using the above lemma, we now obtain the following result.

THEOREM 5.4. *Let $\gamma > 0$ and $P \in \mathbb{R}^{n\times m}$. The following two statements are equivalent:*

(i) *P is a symmetric solution to the quadratic matrix inequality at γ such that* rank $F_\gamma(P) =$ normrank G.

(ii) $$P=\begin{pmatrix} P_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where P_{11} is a symmetric matrix satisfying (5.8).

Furthermore, if the above holds, then the following two statements are equivalent:

(iii) $$\operatorname{rank}\begin{pmatrix} L_\gamma(P,s) \\ F_\gamma(P) \end{pmatrix} = n + \text{normrank } G \text{ for all } s \in \mathbb{C}^0 \cup \mathbb{C}^+.$$

(iv) $$\sigma(A_{11}+\gamma^{-2}E_1E_1^TP_{11}-B_{11}(D_1^TD_1)^{-1}B_{11}^TP_{11} - A_{13}(C_{23}^TC_{23})^{-1}(A_{13}^TP_{11}+C_{23}^TC_{21})) \subset \mathbb{C}^-.$$

Proof. By (5.10) we have $M_\gamma(P) \geqq 0$ if and only if $F_\gamma(P) \geqq 0$, and we also know that these matrices have the same rank. Assume a symmetric P satisfies $M_\gamma(P) \geq 0$ and rank $M_\gamma(P) =$ normrank G. Since $P\mathcal{T}(\Sigma) = 0$ (see Lemma 5.3) we know that we can write P as

$$P=\begin{pmatrix} P_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.12}$$

If we also use the decompositions (4.2) for the other matrices, then we find that $M_\gamma(P)$ is equal to

$$\begin{pmatrix} P_{11}A_{11}+A_{11}^TP_{11}+C_{21}^TC_{21}+\gamma^{-2}P_{11}E_1E_1^TP_{11} & 0 & P_{11}A_{13}+C_{21}^TC_{23} & P_{11}B_{11} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ A_{13}^TP_{11}+C_{23}^TC_{21} & 0 & \boxed{C_{23}^TC_{23}} & 0 & 0 \\ B_{11}^TP_{11} & 0 & 0 & \boxed{D_1^TD_1} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \geqq 0. \tag{5.13}$$

According to Lemma 4.4 the rank of this matrix equals the rank of the encircled matrix. Thus, the Schur complement of the encircled matrix must be equal to zero. Since this condition exactly yields the algebraic Riccati equation (5.8) we find that P_{11} is a solution of (5.8).

Conversely, if P_{11} is a solution of (5.8), then the Schur complement of the encircled matrix in (5.13) is zero. Therefore, it satisfies the matrix inequality (5.13), and the rank of the matrix is equal to normrank G. Hence P given by (5.12) satisfies the required properties.

Now assume that (i) or (ii) holds. We will prove the equivalence of (iii) and (iv). Denote the matrix in (iv) by Z. We will apply the following unimodular transformation

to the matrix in (iii):

$$\begin{pmatrix} I & 0 & 0 \\ 0 & I & F_0^T \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} L_\gamma(P,s) \\ F_\gamma(P) \end{pmatrix} \begin{pmatrix} I & 0 \\ F_0 & I \end{pmatrix}.$$

When we use the decompositions in (4.2), the latter turns out to be equal to

$$\begin{pmatrix} sI-A_{11}-\gamma^{-2}E_1E_1^TP_{11} & 0 & -A_{13} & -B_{11} & 0 \\ -A_{21}-\gamma^{-2}E_2E_1^TP_{11} & sI-A_{22} & -A_{23} & -B_{21} & -B_{22} \\ -A_{31}-\gamma^{-2}E_3E_1^TP_{11} & -A_{32} & sI-A_{33} & -B_{31} & -B_{32} \\ P_{11}A_{11}+A_{11}^TP_{11}+C_{21}^TC_{21}+\gamma^{-2}P_{11}E_1E_1^TP_{11} & 0 & P_{11}A_{13}+C_{21}^TC_{23} & P_{11}B_{11} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ A_{13}^TP_{11}+C_{23}^TC_{21} & 0 & C_{23}^TC_{23} & 0 & 0 \\ B_{11}^TP_{11} & 0 & 0 & D_1^TD_1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.14}$$

By using Schur complements, we can get the Riccati equation (5.8) in the 4,1 position and the matrix Z in the 1,1 position of the above matrix. Furthermore, since $D_1^TD_1$ is invertible, we can make the 2,4 and 3,4 blocks equal to zero by a unimodular transformation. Since P_{11} is a solution of the Riccati equation, the 4,1 block becomes zero. Thus, we find that (5.14) is unimodularly equivalent to

$$\begin{pmatrix} sI-Z & 0 & 0 & 0 & 0 \\ * & sI-A_{22} & -A_{23} & 0 & -B_{22} \\ * & -A_{32} & sI-A_{33} & 0 & -B_{32} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & C_{23}^TC_{23} & 0 & 0 \\ 0 & 0 & 0 & D_1^TD_1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where $*$ denotes matrices that are unimportant for this argument.

Now by Lemma 4.2 the encircled matrices together form the system matrix of a strongly controllable system. Hence this system matrix is unimodularly equivalent to a constant matrix $(I\ 0)$, where I denotes the identity matrix of appropriate size. Therefore, we can make the 2,1 and 3,1 blocks zero by a unimodular transformation. Thus, after reordering we find

$$\begin{pmatrix} L_\gamma(P,s) \\ F_\gamma(P) \end{pmatrix} \sim \begin{pmatrix} sI-Z & 0 & 0 & 0 & 0 \\ 0 & sI-A_{22} & -A_{23} & -B_{22} & 0 \\ 0 & -A_{32} & sI-A_{33} & -B_{32} & 0 \\ 0 & 0 & C_{23}^TC_{23} & 0 & 0 \\ 0 & 0 & 0 & 0 & D_1^TD_1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

It follows that the matrix on the left has rank $n+\text{normrank } G$ for all $s \in \mathbb{C}^0 \cup \mathbb{C}^+$ if and only if $\sigma(Z) \subset \mathbb{C}^-$. This proves that (iii) and (iv) are equivalent. □

A proof of the implication (i)⇒(ii) in Theorem 2.1 is now obtained immediately by combining Corollary 5.2 and Theorem 5.4.

6. Existence of state feedback control laws. In this section we give a proof of the implication (ii)⇒(i) in Theorem 2.1. We first explain the idea of the proof. Again, we consider our control system (5.3) as the interconnection of the subsystem $\tilde{\Sigma}$ given by (5.4), (5.6) and the subsystem Σ_0 given by (5.5). Suppose that the quadratic matrix inequality has a positive-semidefinite solution at γ such that the rank conditions (2.5) and (2.6) hold. Then according to Theorem 5.4, the algebraic Riccati equation associated with the subsystem $\tilde{\Sigma}$ has a positive-semidefinite solution P_{11} such that (iv) of Theorem 5.4 holds. Thus by applying Theorem 5.1 to the subsystem $\tilde{\Sigma}$, we find that the "feedback law"

$$v_1 = -(D_1^T D_1)^{-1} B_{11}^T P_{11} x_1, \tag{6.1}$$

$$x_3 = -(C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}) x_1, \tag{6.2}$$

yields a closed-loop transfer matrix for $\tilde{\Sigma}$ with H_∞ norm smaller than γ. Now we will do the following: construct a state feedback law for the *original* system (5.3) in such a way that in the subsystem $\tilde{\Sigma}$ the equality (6.2) holds *approximately*. The closed-loop transfer matrix of the original system will then be approximately equal to that of the subsystem $\tilde{\Sigma}$ and will therefore also have H_∞ norm smaller than γ.

In our proof an important role will be played by a result in the context of the problem of *almost disturbance decoupling* as studied in [19] and [22]. We will first recall this result here. For the moment assume that we have the following system:

$$\dot{x} = Ax + Bu + Ew, \qquad z = Cx. \tag{6.3}$$

For this system, the almost disturbance decoupling problem with pole placement (ADDPPP) is formulated as follows. For all $\varepsilon > 0$ and for all $M \in \mathbb{R}$, find $F \in \mathbb{R}^{m \times n}$ such that $\|G_F\|_\infty < \varepsilon$ and $\sigma(A+BF) \subset \{s \in \mathbb{C} \mid \operatorname{Re} s < M\}$. It is shown in [19] and [22] that conditions for the existence of such F can be stated in terms of the strongly controllable subspace $\mathcal{T}(\Sigma)$ associated with the system $\Sigma = (A, B, C, 0)$. (In fact, in [19] and [22] this subspace is denoted by $\mathcal{R}_b^*(\ker C)$.) The exact result is as follows.

LEMMA 6.1. *Consider the system* (6.3). *Let* $\mathcal{T}(\Sigma)$ *denote the strongly controllable subspace associated with* $\Sigma = (A, B, C, 0)$. *Then the following two statements are equivalent*:

(i) *For all* $\varepsilon > 0$ *and for all* $M \in \mathbb{R}$ *there exists* $F \in \mathbb{R}^{m \times n}$ *such that* $\|G_F\|_\infty < \varepsilon$ *and* $\sigma(A+BF) \subset \{s \in \mathbb{C} \mid \operatorname{Re} s < M\}$.

(ii) $\operatorname{im} E \subset \mathcal{T}(\Sigma)$ *and* (A, B) *is controllable.*

As an immediate consequence of the above we obtain the following fact. If $\Sigma = (A, B, C, 0)$ is strongly controllable, then for all $\varepsilon > 0$ and for all $M \in \mathbb{R}$ there exists $F \in \mathbb{R}^{m \times n}$ such that $\|G_F\|_\infty < \varepsilon$ and $\sigma(A+BF) \subset \{s \in \mathbb{C} \mid \operatorname{Re} s < M\}$. Thus, in particular, if $\Sigma = (A, B, C, 0)$ is strongly controllable, then for all $\varepsilon > 0$ there exists $F \in \mathbb{R}^{m \times n}$ such that $\|G_F\|_\infty < \varepsilon$ and $\sigma(A+BF) \subset \mathbb{C}^-$.

We now formulate and prove the converse of Corollary 5.2.

THEOREM 6.2. *Consider the system* (2.1). *Assume that* (A, B, C, D) *has no invariant zeros in* $\mathbb{C}^0$. *Let* $\gamma > 0$. *Assume there exists a real symmetric solution* $P_{11} \geqq 0$ *to the algebraic Riccati equation* (5.8) *such that* (5.9) *holds. Then there exists* $F \in \mathbb{R}^{m \times n}$ *such that* $\|G_F\|_\infty < \gamma$ *and* $\sigma(A+BF) \subset \mathbb{C}^-$.

Proof. Clearly, it is sufficient to prove the existence of such a state feedback law as $v = Fx$ for the system (5.3). Let this system be decomposed according to (5.4)-(5.6). Choose

$$v_1 = -(D_1^T D_1)^{-1} B_{11}^T P_{11} x_1$$

and introduce a new state variable q_3 by

$$q_3 := x_3 + (C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}) x_1 .$$

Then (5.4)-(5.6) can be rewritten as

$$\dot{x}_1 = \tilde{A}_{11} x_1 + A_{13} q_3 + E_1 w, \tag{6.4}$$

$$\begin{pmatrix} \dot{x}_2 \\ \dot{q}_3 \end{pmatrix} = \begin{pmatrix} A_{22} & A_{23} \\ A_{32} & \tilde{A}_{33} \end{pmatrix} \begin{pmatrix} x_2 \\ q_3 \end{pmatrix} + \begin{pmatrix} B_{22} \\ B_{32} \end{pmatrix} v_2 + \begin{pmatrix} \tilde{A}_{21} \\ \tilde{A}_{31} \end{pmatrix} x_1 + \begin{pmatrix} E_2 \\ \tilde{E}_3 \end{pmatrix} w, \tag{6.5}$$

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \tilde{C}_1 \\ \tilde{C}_2 \end{pmatrix} x_1 + \begin{pmatrix} 0 \\ C_{23} \end{pmatrix} q_3 . \tag{6.6}$$

Here we use the following definitions:

$$\begin{aligned}
\tilde{A}_{11} &:= A_{11} - A_{13}(C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}) - B_{11}(D_1^T D_1)^{-1} B_{11}^T P_{11}, \\
\tilde{A}_{21} &:= A_{21} - A_{23}(C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}) - B_{21}(D_1^T D_1)^{-1} B_{11}^T P_{11}, \\
\tilde{A}_{31} &:= A_{31} - A_{33}(C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}) - B_{31}(D_1^T D_1)^{-1} B_{11}^T P_{11} \\
&\quad + (C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}) \tilde{A}_{11}, \\
\tilde{A}_{33} &:= A_{33} + (C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}) A_{13}, \\
\tilde{C}_1 &:= -D_1(D_1^T D_1)^{-1} B_{11}^T P_{11}, \\
\tilde{C}_2 &:= C_{21} - C_{23}(C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}), \\
\tilde{E}_3 &:= E_3 + (C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}) E_1 .
\end{aligned}$$

According to Theorem 5.1, if in the subsystem formed by (6.4) and (6.6) we have $q_3 = 0$, then its transfer matrix from w to z has H_∞ norm smaller than γ. Moreover, we have $\sigma(\tilde{A}_{11}) \subset \mathbb{C}^-$. Hence, there exist $M > 0$ and $\rho > 0$ such that for all w and q_3 in $\mathscr{L}_2$, we have

$$\|z\|_2 < (\gamma - \rho)\|w\|_2 + M\|q_3\|_2 . \tag{6.7}$$

Also by the fact that $\tilde{A}_{11}$ is stable, there exist $M_1, M_2 > 0$ such that for all w and q_3 in $\mathscr{L}_2$, we have

$$\|x_1\|_2 < M_1\|w\|_2 + M_2\|q_3\|_2 . \tag{6.8}$$

We claim that the following system is strongly controllable:

$$\left(\begin{pmatrix} A_{22} & A_{23} \\ A_{32} & \tilde{A}_{33} \end{pmatrix}, \begin{pmatrix} B_{22} \\ B_{32} \end{pmatrix}, (0 \quad I), 0 \right). \tag{6.9}$$

This can be seen by the following transformation:

$$\left(\begin{array}{ccc} I & 0 & 0 \\ 0 & I & A_{33}-\tilde{A}_{33} \\ 0 & 0 & I \end{array}\right)\left(\begin{array}{cc:c} sI-A_{22} & -A_{23} & -B_{22} \\ -A_{32} & sI-A_{33} & -B_{32} \\ \hdashline 0 & I & 0 \end{array}\right)=\left(\begin{array}{cc:c} sI-A_{22} & -A_{23} & -B_{22} \\ -A_{32} & sI-\tilde{A}_{33} & -B_{32} \\ \hdashline 0 & I & 0 \end{array}\right).$$

Since the first matrix on the left is unimodular and the second matrix has full row rank for all $s \in \mathbb{C}$ (see Lemma 4.2), the matrix on the right has full row rank for all $s \in \mathbb{C}$. Hence the system (6.9) is strongly controllable.

Now consider the almost disturbance decoupling problem for the system (6.5) with output q_3 and "disturbance" $\binom{x_1}{w}$. Because of strong controllability of (6.9) there exists a feedback law $v_2 = F_1\binom{x_2}{q_3}$ such that in (6.5) we have

$$\|q_3\|_2 < \tfrac{1}{2}\rho(M + M_1 M + \rho M_2)^{-1}\{\|w\|_2 + \|x_1\|_2\} \tag{6.10}$$

for all w and x_1 in $\mathscr{L}_2$ and such that the matrix

$$\tilde{A} := \begin{pmatrix} A_{22} & A_{23} \\ A_{32} & \tilde{A}_{33} \end{pmatrix} + \begin{pmatrix} B_{22} \\ B_{32} \end{pmatrix} F_1$$

satisfies $\sigma(\tilde{A}) \subset \mathbb{C}^-$. Combining (6.7), (6.8), and (6.10) gives us

$$\|z\|_2 < (\gamma - \tfrac{1}{2}\rho)\|w\|_2$$

for all w in $\mathscr{L}_2$. Summarizing, we have now shown that if in our original system (5.3) we apply the state feedback law

$$\begin{aligned} v_1 &= -(D_1^T D_1)^{-1} B_{11}^T P_1 x_1, \\ v_2 &= F_1 \begin{pmatrix} x_2 \\ x_3 + (C_{23}^T C_{23})^{-1}(A_{13}^T P_{11} + C_{23}^T C_{21}) x_1 \end{pmatrix}, \end{aligned} \tag{6.11}$$

then for all $w \in \mathscr{L}_2^l(\mathbb{R}^+)$ we have $\|z\|_2 < \gamma\|w\|_2$. Thus, the H_∞ norm of the resulting closed-loop transfer matrix is smaller than γ.

It remains to be shown that the closed-loop system is internally stable. We know that

$$\|(sI - \tilde{A}_{11})^{-1} A_{13}\|_\infty \leqq M_2, \tag{6.12}$$

$$\left\| (0 \quad I)(sI - \tilde{A})^{-1} \begin{pmatrix} \tilde{A}_{21} \\ \tilde{A}_{31} \end{pmatrix} \right\|_\infty \leqq \tfrac{1}{2}\rho(M + M_1 M + \rho M_2)^{-1} \leqq \tfrac{1}{2} M_2^{-1}. \tag{6.13}$$

The closed-loop A-matrix resulting from the feedback law (6.11) is given by

$$A_{\mathrm{cl}} := \begin{pmatrix} \tilde{A}_{11} & (0 \quad A_{13}) \\ \begin{pmatrix} \tilde{A}_{21} \\ \tilde{A}_{31} \end{pmatrix} & \tilde{A} \end{pmatrix}.$$

Assume $(x^T, y^T, z^T)^T$ is an eigenvector of A_{cl} with eigenvalue λ with $\operatorname{Re}\lambda \geqq 0$. It can be seen that

$$x = (\lambda I - \tilde{A}_{11})^{-1} A_{13} z, \tag{6.14}$$

$$z = (0 \quad I)(\lambda I - \tilde{A})^{-1} \begin{pmatrix} \tilde{A}_{21} \\ \tilde{A}_{31} \end{pmatrix} x. \tag{6.15}$$

(Note that the inverses exist due to the fact that $\tilde{A}_{11}$ and $\tilde{A}$ are stable matrices.)

Combining (6.12) and (6.14) we find $\|x\| \leqq M_2\|z\|$, and combining (6.13) and (6.15) yields $\|z\|_2 \leqq \frac{1}{2}M_2^{-1}\|x\|_2$. Hence $x = z = 0$. This, however, would imply that $(y^T \ 0^T)^T$ is an unstable eigenvector of $\tilde{A}$. Since $\sigma(\tilde{A}) \subset \mathbb{C}^-$, this yields a contradiction. This proves that the closed-loop system is internally stable. □

A proof of the implication (ii)⇒(i) in Theorem 2.1 is now obtained by combining Theorems 5.4 and 6.2.

Remark 6.3. In the regular case (i.e., D injective) it is quite easy to give an explicit expression for a suitable state feedback law. Indeed, if $P \geqq 0$ is a solution to the algebraic Riccati equation (5.1) such that (5.2) holds, then the feedback law $u = -(D^TD)^{-1}(B^TP + D^TC)x$ achieves internal stability and $\|G_F\|_\infty < \gamma$. In the singular case (i.e., D not injective) a state feedback law is given by $u = F_0x + v$. Here, F_0 is given by (4.1) and $v = (v_1^T, v_2^T)^T$ is given by (6.11). The matrix P_{11} is obtained by solving the quadratic matrix inequality or, equivalently, by solving the reduced order Riccati equation (5.8). The matrix F_1 is a "state feedback" for the strongly controllable auxiliary system (6.5). This state feedback achieves almost disturbance decoupling between the "disturbance" $(x_1^T, w^T)^T$ and the "output" q_3. The required accuracy of decoupling is expressed by (6.10). A conceptual algorithm to construct such F_1 can be based on the proof of [19, Thm. 3.36].

7. Discussion and conclusions. In this paper we have shown that if in the H_∞ problem with state feedback *no* assumptions are made on the direct feedthrough matrix of the control input, then the central role of the algebraic Riccati equation is taken over by a quadratic matrix inequality. We note that a similar phenomenon is known to occur in the linear quadratic regulator problem: if the weighting matrix of the control input is singular, then the optimal cost is given in terms of a (linear) matrix inequality rather than in terms of an algebraic Riccati equation (see [21]). However, while in the singular LQ problem optimal inputs in general are distributions, in the H_∞ context *also in the singular case suitable state feedback laws can be found.* It is well known that in the LQ problem a special role is played by solutions of the linear matrix inequality that *minimize* the rank of the dissipation matrix (see [4], [13]). It turns out that also in our context the relevant solutions to the quadratic matrix inequality are *rank minimizing.* Indeed, it follows from the proof of Theorem 5.4 that for *all* symmetric matrices P we have rank $F_\gamma(P) \geqq$ normrank G. Thus, (2.5) can be interpreted as saying that P minimizes the rank of $F_\gamma(P)$. On the other hand, once we know that rank $F_\gamma(P) =$ normrank G, then obviously for all $s \in \mathbb{C}$ we have

$$\operatorname{rank}\begin{pmatrix} L_\gamma(P, s) \\ F_\gamma(P) \end{pmatrix} \leqq n + \text{normrank } G.$$

Thus, (ii) of Theorem 2.1 can, loosely speaking, be reformulated as follows. There exists a solution $P \geqq 0$ to $F_\gamma(P) \geqq 0$ that *minimizes* rank $F_\gamma(P)$ and *maximizes* rank $(L_\gamma(P, s)^T, F_\gamma(P)^T)^T$ for all $s \in \mathbb{C}^0 \cup \mathbb{C}^+$.

As can be expected, the quadratic matrix inequality and the rank conditions (2.5) and (2.6) turn out to play an important role in the context of *singular linear quadratic differential games.* This connection is elaborated in [16].

Needless to say, several questions remain unanswered in this paper. The most obvious topic is the extension of the theory of this paper to the case of dynamic measurement feedback, i.e., the singular counterpart of the problem studied in [2], [5], and [18]. In [17] it is shown that the existence of suitable dynamic compensators require solvability of a *pair* of quadratic matrix inequalities.

Finally, in [20] the ideas of the present paper are used to tackle the *finite horizon* "H_∞" control problem by measurement feedback, i.e., the problem of finding a dynamic

compensator such that the $L_2[t_0, t_1]$-induced norm (instead of the $L_2(\mathbb{R}^+)$-induced norm) of the closed-loop operator is smaller than an a priori given upper bound. In [20] conditions for the existence of such a compensator are formulated in terms of quadratic differential inequalities (the extensions of Riccati differential equations).

REFERENCES

[1] J. A. BALL AND N. COHEN, *Sensitivity minimization in the H^∞ norm: parametrization of all suboptimal solutions*, Internat. J. Control, 46 (1987), pp. 785-816.

[2] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831-847.

[3] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1959.

[4] T. GEERTS, *All optimal controls for the singular linear-quadratic problem without stability; a new interpretation of the optimal cost*, Linear Algebra Appl., 122 (1989), pp. 65-104.

[5] K. GLOVER AND J. C. DOYLE, *State space formulae for all stabilizing controllers that satisfy an H_∞ norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167-172.

[6] M. L. J. HAUTUS, *Strong detectability and observers*, Linear Algebra Appl., 50 (1983), pp. 353-368.

[7] M. L. J. HAUTUS AND L. M. SILVERMAN, *System structure and singular control*, Linear Algebra Appl., 50 (1983), pp. 369-402.

[8] P. P. KHARGONEKAR, I. R. PETERSEN, AND M. A. ROTEA, *H_∞ optimal control with state feedback*, IEEE Trans. Automat. Control, 33 (1988), pp. 786-788.

[9] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[10] I. R. PETERSEN, *Disturbance attenuation and H^∞ optimization: a design method based on the algebraic Riccati equation*, IEEE Trans. Automat. Control, 32 (1987), pp. 427-429.

[11] H. H. ROSENBROCK, *The zeros of a system*, Internat. J. Control, 18 (1973), pp. 297-299.

[12] J. M. SCHUMACHER, *Dynamic Feedback in Finite and Infinite Dimensional Linear Systems*, Mathematical Centre Tracts, Vol. 143, Amsterdam, 1981.

[13] ———, *The role of the dissipation matrix in singular optimal control*, System Control Lett., 2 (1983), pp. 262-266.

[14] ———, *On the structure of strongly controllable systems*, Internat. J. Control, 38 (1983), pp. 525-545.

[15] A. A. STOORVOGEL, *H_∞ control with state feedback*, Proceedings MTNS-89, Amsterdam, 1990.

[16] ———, *The singular zero-sum differential game with stability using H_∞ control theory*, Math. Control Sign. Systems, to appear.

[17] ———, *The singular H_∞ control problem with dynamic measurement feedback*, SIAM J. Control Optim., to appear.

[18] G. TADMOR, *H_∞ in the time domain: the standard four blocks problem*, Math. Control Sign. Systems, to appear.

[19] H. L. TRENTELMAN, *Almost Invariant Subspaces and High Gain Feedback*, CWI Tracts, Vol. 29, Amsterdam, 1986.

[20] H. L. TRENTELMAN AND A. A. STOORVOGEL, *Completion of the squares in the finite horizon H^∞ control problem by measurement feedback*, preprint, October 1989.

[21] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621-634.

[22] ———, *Almost invariant subspaces: an approach to high gain feedback design—Part 1: almost controlled invariant subspaces*, IEEE Trans. Automat. Control, 26 (1981), pp. 235-252.

[23] K. ZHOU AND P. P. KHARGONEKAR, *An algebraic Riccati equation approach to H^∞ optimization*, Systems Control Lett., 11 (1988), pp. 85-91.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 1209-1233, September 1990

010

ROUTING AND SINGULAR CONTROL FOR QUEUEING NETWORKS IN HEAVY TRAFFIC*

LUIZ FELIPE MARTINS† AND HAROLD J. KUSHNER‡

Abstract. The problem of routing control in an open queueing network under conditions of heavy traffic and finite (scaled) buffers is dealt with. The operating statistics can be state dependent. The sequence of scaled controlled state processes converges to a singularly controlled reflected diffusion (with the associated costs), under broad conditions. Due to the nature of the controls, a "scaling" method is introduced to obtain the convergence, since the actual sequence of processes does not necessarily converge in the Skorokhod topology. Owing to finite buffers, an extension of the reflection mapping needs to be obtained. The optimal value functions for the physical processes converge to the optimal value function of the limit process, under broad conditions. Approximations to the optimal control for the limit process are obtained, as well as properties of the sequence of physical processes. The optimal or controlled (but not necessarily optimal) limit process can be used to approximate a large variety of functionals of the optimal or controlled (but not necessarily optimal) physical processes.

Key words. routing control, weak convergence, singular control, queues in heavy traffic, reflected controlled diffusions

AMS(MOS) subject classifications. 90B15, 90B22, 93E20, 93E25, 60F17

1. Introduction.

We consider the problem of optimal or nearly optimal routing in a queueing system under heavy traffic conditions. The general network model is a "controlled routing" form of the general open network dealt with by Reiman [1], where each customer eventually leaves the system. See also Harrison and Reiman [2] and Harrison [10] for a discussion of models that are limits of such systems. We will actually treat two special cases for simplicity in the development. But it should be apparent from these cases that the general open network can be treated in the same way. The treated cases involve all the basic techniques that are required for the general case. In [1], there is a finite set of servers, each with an infinite buffer. We bound (and appropriately scale) the buffers here. It is well known [1], [2], that under broad conditions on the service and interarrival times, the vector of queue length processes (with an appropriate amplitude normalization and time scaling) converges weakly to a reflected diffusion, as the traffic intensity goes to unity.

The work in [1] requires that the system operating statistics not be state dependent, and uses results for the weak convergence of a sequence of sums of mutually independent random variables to a Wiener process, together with a clever method to treat the boundary to get the appropriate limit. The methods that are used to identify the limit as a reflected diffusion are not extendible to the state-dependent or to the controlled case, where the required independence no longer holds, and the characterization of the limit processes, as well as the proofs of tightness, require different methods. The "martingale type" methods for getting limit theorems for wide bandwidth noise driven systems seem to be more appropriate for characterizing the limit process. In [4] there is a study of a heavy traffic problem under a control, and the arrival and service processes are allowed to be state dependent. Such state dependence is natural for the

* Received by the editors April 27, 1989; accepted for publication (in revised form) October 30, 1989.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912 and Mathematics Department of Universidade Federal do Rio Grande do Sul, Brazil.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This work was supported in part by Air Force Office of Scientific Research grant 89-0015 and Army Research Office grant DAAL-03-86-K-0171.

controlled problem, since we might want to let the processing depend on what is happening in the system. In addition, the methods that are needed to characterize the controls in the limit problem as "nonanticipative," etc., require the use of the same methods that the state dependence requires.

In [4], the processors and the arrival sequences can be turned on or off to control the flows and the costs. The limit problem is an impulsively controlled reflected diffusion of a nonclassical type, since there is the possibility of multiple "simultaneous" impulses. It is shown in [4] that any sequence of controlled physical processes with uniformly bounded costs converged to a well-defined controlled limit process. Also, the sequence of optimally controlled physical processes converges to the optimally controlled limit process, in the sense that the value functions converge. Also a control that is nearly optimal for the limit process can be adapted to become a nearly optimal control for the physical process under heavy traffic, under quite broad conditions. Such results help to justify the use of heavy traffic limit theorems for optimal or other control purposes. Because of the behavior of the physical process in [4] when the on-off controls are used, the Skorokhod topology must be used with care, because the actual scaled queue length processes do not converge in the Skorokhod topology as it is usually used. Also, that reference provides convergent numerical algorithms.

In this paper, we also deal with a controlled heavy traffic problem. In the basic model, the routing of a subset of the external arrivals could be controlled. The aims are similar to those in [4]. The dynamical equations for the scaled queue length process are defined. The sequence of such processes (as the traffic intensity tends to unity) might not be tight in the Skorokhod topology, due to the nature of the routing control. To handle this, we start by working with a rescaling of the time, with which we can get tightness, and a characterization of the weak limits. The rescaling depends on the control. After the limits are obtained, an "inverse" scaling (dependent on the limit control) yields the process that actually characterizes the limit of the cost functionals. The limit process is a controlled reflected diffusion. But the control is of the "singular" type in the sense of [8]. The usual reflection mapping that is used to handle the problem of nonnegativity of the queue length process must be modified here, due to the presence of the finite buffer. We construct the proper reflection mapping from a sequence of concatenations of the usual one.

The basic problem of interest is defined in § 2. For notational simplicity we work with a system of only two processors. Also, until § 7, we do not have feedback. The addition of feedback is straightforward, but it seems to be preferable to present the ideas in as unencumbered a fashion as possible. The extension of the result to the general routing controlled open network is straightforward. Some of the weak convergence arguments and definitions from [4] are used, but familiarity with that reference is not necessary. In § 2, we manipulate the state equations into the "martingale plus drift" form that will be used in the weak convergence arguments. The reflection mapping result is stated in § 3 (and proved in § 8). The required rescaling is defined and the tightness and weak convergence proved in § 3. We must prove that the limit (singular) controls are nonanticipative with respect to the Wiener processes, which "drive" the limit process.

Section 4 is concerned with the convergence of the cost functions. We prove that there is a routing control with a uniformly bounded cost, and show that the liminf of the optimal cost functions for the physical processes is bounded below by the optimal cost for the limit process. To show that the limit of the optimal costs for the physical processes is the optimal cost for the limit process, we need to prove various existence and approximation results for the optimal policy for the limit problem. This is done

in § 6, and uses the "limit form" of the control-dependent rescaling introduced in § 3. An interesting approach to the approximation problem is discussed. The general rescaling and tightness methods are of much wider use for limit and approximation problems where singular controls are involved and where there might not be convergence in the Skorokhod topology. The developed approximations are then used to prove the approximate optimality for the physical processes of an appropriate nearly optimal policy for the limit process.

Approximations to singular control problems for wide bandwidth noise driven systems have been discussed in [6], but the method used here is rather different and is very natural for the kinds of problems that are being considered. Numerical methods have been developed for the problems of this paper. The proofs of their convergence require methods that are similar to those used here, but since there are many additional details, they will be dealt with in a subsequent paper.

There has been considerable work done in controlled routing, including some formal work on routing under heavy traffic [11]. For the types of problems considered here, or for reasonable extensions, it seems to be nearly impossible to obtain the optimal or nearly optimal strategies. The idea here is to use the relative simplicity of heavy traffic limits to get an optimal control problem that can be solved numerically, and then to use an appropriate adaptation of that solution for the true physical problem. The methods are applicable to a wide variety of problems.

The problem in [4] has an impulsively controlled limit, since the costs associated with the control actions are bounded away from zero. In the present case, the troublesome part of the cost is the "scaled number of customers" that are rerouted. This could lead to an impulsively controlled limit system. But, in general, the limit has a "singular" control component. Consider, e.g., the case where we reroute to one processor if the buffer of another is half-full. Then the limit control will have the structure of a local time at the "half-full" boundary.

2. Problem description. Until § 7, we work with the simple system of Fig. 1. This will enable us to develop the main ideas without an excessive notational burden. Also, for notational convenience, we work with a discrete-time parameter. The results for the analogous continuous-time parameter case are the same. Each of the processors P_0, P_1, and P_2 has its own stream of arrivals from the exterior. P_0 is used only as an (instantaneous) routing node. Its service time is zero. This can readily be changed, and the resulting network would then be a special case of the general network discussed in § 7. The P_0 routes to either P_1 or P_2, and (until § 7), the completed services from P_1 and P_2 leave the system. The routing decision is based on the events up to the time

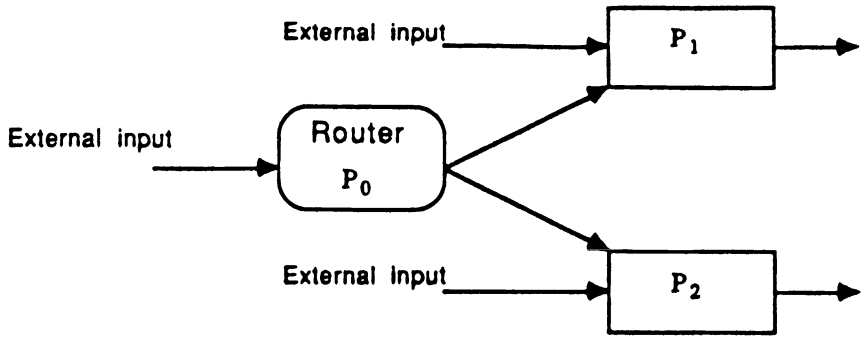


FIG. 1. *A simple routing problem.*

of the decision. We suppose that some prior routing is assigned to each new arrival to P_0, but that the routing node can reassign, with an associated profit or loss. Next, we give some simple examples.

Example 1. There are two classes of customers arriving (at random) at P_0. P_i is more efficient for class i, and a prior assignment of class i to P_i is made. But P_0 can reroute to the less efficient processor, depending on the system state. The cost of rerouting might be, for example, a setup cost, or the *relative* cost of the less efficient processor.

Example 2. The case of Example 1, but with three classes of customers, arriving at random. Class i ($i=1$ or 2) must be served by P_i. Class 3 can be served by either processor, but one of the P_i is more efficient (cheaper) and a prior assignment to that P_i is made. But P_0 can alter the assignment. For example, let the P_i represent data bases, with some overlap of data files. A subset of the arriving jobs needs only the "overlap" data. But one of the P_i is "faster" than the other.

Example 3. P_1 is cheaper for all customers arriving at P_0. But, due to the heavy traffic conditions, the mean number of customers routed to each P_i is essentially fixed (modulo some fraction that goes to zero as the traffic intensity goes to unity). Some prior assignment is made, but P_0 can reroute at either a cost or a savings if appropriate.

In most of the development, we let the distribution of the processing time depend only on the processor and not on the customer type. The general case is a minor extension and is discussed at the end of § 3. In general, the model can be readily extended to handle rerouting of a customer actually in a queue as well as reneging.

In the modeling of systems under heavy traffic conditions, it has been the usual practice to suppose that the processors "keep processing" and create departures even if the queues are empty [1]–[4]. Whatever "fictitious" departures occur due to this convention are compensated for by an added "reflection term" (our Y below). Thus each P_i ($i=1,2$) has associated to it a sequence of service intervals that cover all time. This convention simplifies the analysis. Also, we suppose (as is the usual practice) that if a customer arrives at P_1 or P_2 when the associated queue is empty, then the service time for that customer is just the residual time of the current service time interval for that processor. As in [1]–[4], this convention does not affect the limit processes.

It is possible that multiple events can occur at the same time at P_1 or P_2. This could happen even if we worked in continuous time. For the sake of precision, we suppose that a departure (real or fictitious) from a processor always occurs "just before" any arrival to that processor, and that if two arrivals to the same P_i occur at the same time, then the one from P_0 takes precedence. Such a conflict might arise if there is only space for one customer left in some buffer, but there are two arrivals. We ignore these distinctions in the notation, for simplicity. It can be shown that the precedence relations do not affect the limit.

DEFINITIONS. We use the notation of [4] whenever possible, although knowledge of that reference is not needed for the reading of this paper. The symbol ε indexes the traffic intensity; as $\varepsilon \to 0$, the intensity goes to one. For each $\varepsilon > 0$ and $i = 1, 2$, let $\{\Delta_n^{i,\varepsilon}, n = 1, 2, \cdots\}$, denote the sequence of service times for P_i and let $\psi_n^{i,\varepsilon}$ be the indicator function of the event that a service (real or fictitious) is completed at P_i at time n. For $i = 0, 1, 2$ and each $\varepsilon > 0$, let $\{\alpha_n^{i,\varepsilon}, n < \infty\}$ denote the sequence of interarrival times to P_i, from the exterior of the system, and let $\xi_n^{i,\varepsilon}$, $i = 0, 1, 2$, be the indicator of the event that there is an external arrival to P_i at time n. Write t/ε for $[t/\varepsilon]$, the largest integer, which is no bigger than t/ε. Define $X_n^{i,\varepsilon} = \sqrt{\varepsilon}$ (number of customers waiting for or in service at P_i at time n), and set $X^{i,\varepsilon}(t) = X_{t/\varepsilon}^{i,\varepsilon}$. In general, for a sequence

$\{Z_n^\varepsilon\}$, define the function $Z^\varepsilon(t) = Z^\varepsilon_{t/\varepsilon}$. The buffer of P_i, $i = 1, 2$, has size $B_i/\sqrt{\varepsilon}$, which we assume is always an integer. Let $I_n^{i,\varepsilon}$ denote the indicator of the event that an arrival at P_0 at time n has the *prior* assignment to P_i, and let $\rho_n^{ij,\varepsilon}$, $j \neq i$ be the indicator of the event that this arrival is *reassigned* to P_j.

We usually use the convention that the superscript ε is dropped whenever one of the above terms is used as a summand. The notation would not be much simpler if we worked in continuous time, since we would still have to keep track of the events and their times. Define $(j \neq i)$

$$A_n^{i,\varepsilon} = \sqrt{\varepsilon} \sum_{m=0}^{n} \xi_m^i, \qquad A_n^{0i,\varepsilon} = \sqrt{\varepsilon} \sum_{m=0}^{n} I_m^i \xi_m^0,$$

$$D_n^{i,\varepsilon} = \sqrt{\varepsilon} \sum_{m=0}^{n} \psi_m^i, \qquad J_n^{ij,\varepsilon} = \sqrt{\varepsilon} \sum_{m=0}^{n} \xi_m^0 I_m^i \rho_m^{ij},$$

$$J_n^{i,\varepsilon} = J_n^{ji,\varepsilon} - J_n^{ij,\varepsilon}, \qquad Y_n^{i,\varepsilon} = \sqrt{\varepsilon} \sum_{m=0}^{n} \psi_m^i I_{\{X_m^i = 0\}},$$

$$U_n^{i,\varepsilon} = \sqrt{\varepsilon} \sum_{m=0}^{n} \xi_m^i I_{\{X_m^i = B_i\}} + \sqrt{\varepsilon} \sum_{m=0}^{n} \xi_m^0 (I_m^i + I_m^j \rho_m^{ji} - I_m^i \rho_m^{ij}) I_{\{X_m^i = B_i\}}.$$

The $A_n^{0i,\varepsilon}$ is the scaled total number of arrivals (by time n) at P_0 that have been a priori assigned to P_i (they might, of course, be rerouted by P_0).

The $J_n^{ij,\varepsilon}$ are the "rerouting" control terms, the scaled number of customers originally destined for P_i but rerouted to P_j. The $Y_n^{i,\varepsilon}$ is the scaled total number of "fictitious" departures due to our convention of continuing to "process" even if the queue is empty, and $U_n^{i,\varepsilon}$ is the number of customers lost to P_i when its buffer is full.

The *mass balance* equations can be written as (discrete "real" and "interpolated" time, respectively),

$$X_n^{i,\varepsilon} = X_0^{i,\varepsilon} + A_n^{i,\varepsilon} + A_n^{0i,\varepsilon} - D_n^{i,\varepsilon} + J_n^{i,\varepsilon} + Y_n^{i,\varepsilon} - U_n^{i,\varepsilon}, \tag{2.1}$$

$$X^{i,\varepsilon}(t) = X_0^{i,\varepsilon} + A^{i,\varepsilon}(t) + A^{0i,\varepsilon}(t) - D^{i,\varepsilon}(t) + J^{i,\varepsilon}(t) + Y^{i,\varepsilon}(t) - U^{i,\varepsilon}(t). \tag{2.2}$$

The cost function. Let $\beta > 0$, $c_i > 0$, $k_i > 0$, and let $k(\cdot)$ be a bounded and continuous function. Define $J^\varepsilon = (J^{12,\varepsilon}, J^{21,\varepsilon})$. We use the cost functional

$$\begin{aligned} V^\varepsilon(x, J^\varepsilon) = E_x \int_0^\infty e^{-\beta t} k(X^\varepsilon(t))\, dt + E_x \int_0^\infty e^{-\beta t} [k_1\, dJ^{12,\varepsilon}(t) \\ + k_2\, dJ^{21,\varepsilon}(t) + c_1\, dU^{1,\varepsilon}(t) + c_2\, dU^{2,\varepsilon}(t)]. \end{aligned} \tag{2.3}$$

By Theorem 7 below, there are routing policies $J^{ij,\varepsilon}(\cdot)$ for which

$$\sup_\varepsilon V^\varepsilon(x, J^\varepsilon) < \infty. \tag{2.4}$$

Define

$$V^\varepsilon(x) = \inf_{J^\varepsilon} V^\varepsilon(x, J^\varepsilon).$$

The $k(\cdot)$ might be nonlinear. Such nonlinear $k(\cdot)$ occur when we wish to model the costs of reneging or queue switching, or if we wish to limit the possibility of leaving the queue due to a "long" wait. The second term in (2.3) penalizes the overflows and rerouting. One of the k_i can be negative and we return to this case at the end of § 6.

Definitions and heavy traffic assumptions. We take many of the definitions from [4] so that the results of that reference can be conveniently used. Define $S^{i,\varepsilon}_{a,n} = \sum_{m=1}^{n} \alpha^i_m$, $S^{i,\varepsilon}_{d,n} = \sum_{m=1}^{n} \Delta^i_m$. Define $\bar{S}^{i,\varepsilon}_a(\cdot)$ by $\bar{S}^{i,\varepsilon}_a(t) = \max\{\varepsilon m: \varepsilon S^{i,\varepsilon}_{a,m} \leqq t\}$, and define $\bar{S}^{i,\varepsilon}_d(\cdot)$ analogously. Actually, $\bar{S}^{i,\varepsilon}_a(t) = \sqrt{\varepsilon}\, A^{i,\varepsilon}(t)$ and analogously for $\bar{S}^{i,\varepsilon}_d(t)$, but the separate terminology is useful. These functions are the "inverses" of the functions $\varepsilon S^{i,\varepsilon}_\alpha(\cdot)$. Let $E^{i,\varepsilon}_{a,n}$ denote the expectation, conditioned on the arrival and departure intervals that started by $S^{i,\varepsilon}_{a,n}$ (except for $\alpha^{i,\varepsilon}_{n+1}$), and the control (routing) actions taken up to $S^{i,\varepsilon}_{a,n}$. Define $E^{i,\varepsilon}_{d,n}$ analogously, where $S^{i,\varepsilon}_{d,n}$ and $\Delta^{i,\varepsilon}_{n+1}$ replace $S^{i,\varepsilon}_{a,n}$ and $\alpha^{i,\varepsilon}_{n+1}$, respectively. Similarly, define the conditional variances $\mathrm{var}^{i,\varepsilon}_{\alpha,n}$, $\alpha = a, d$. We use the notation

$$E^{i,\varepsilon}_{a,n}\alpha^{i,\varepsilon}_{n+1} = \bar{\alpha}^{i,\varepsilon}_{n+1}, \qquad E^{i,\varepsilon}_{d,n}\Delta^{i,\varepsilon}_{n+1} = \bar{\Delta}^{i,\varepsilon}_{n+1},$$

$$\mathrm{var}^{i,\varepsilon}_{a,n}\alpha^{i,\varepsilon}_{n+1} = (\sigma^{i,\varepsilon}_{a,n+1})^2, \qquad \mathrm{var}^{i,\varepsilon}_{d,n}\Delta^{i,\varepsilon}_{n+1} = (\sigma^{i,\varepsilon}_{d,n+1})^2.$$

We will use the following assumptions. Assumptions (A2.1) and (A2.4) are the "usual" heavy traffic assumptions. Assumption (A2.4) basically says that (modulo a term that goes to zero as $\varepsilon \to 0$) the mean rate of arrivals to P_i equals the "capacity" of P_i.

(A2.1) There are real $g_{ai} > 0$, $g_{di} > 0$ and bounded and continuous real-valued functions $a^i(\cdot)$ and $d^i(\cdot)$ such that

$$(\bar{\alpha}^{i,\varepsilon}_{n+1})^{-1} = g_{ai} + \sqrt{\varepsilon}\, a_{in} + o(\sqrt{\varepsilon}),$$

$$(\bar{\Delta}^{i,\varepsilon}_{n+1})^{-1} = g_{di} + \sqrt{\varepsilon}\, d_{in} + o(\sqrt{\varepsilon}),$$

where

$$a_{in} = a^i(X^{\varepsilon}_{S^{i,\varepsilon}_{a,n}}), \qquad d_{in} = d^i(X^{\varepsilon}_{S^{i,\varepsilon}_{d,n}}).$$

Note that $X^{\varepsilon}_{S^{i,\varepsilon}_{d,n}}$ is the value of the state at the beginning of the $(n+1)$st interarrival interval, and so it is the correct argument of the $a^i(\cdot)$ above, and similarly for the $d^i(\cdot)$.

(A2.2) $\{|\alpha^{i,\varepsilon}_n|^2, |\Delta^{i,\varepsilon}_n|^2, i, n, \varepsilon > 0\}$ is uniformly integrable.

(A2.3) There are $\bar{p}_i$ such that $P\{I^{i,\varepsilon}_n = 1 \,|\, \text{all arrival or departure intervals starting by time } n \text{ and routing actions up to time } n-1\} = \bar{p}_i$.

This assumption can be weakened in many ways, allowing for batch rerouting and other variations, as well as correlated routings. All that is really needed is that "loosely speaking," $\bar{p}_i$ be a "local mean" of the conditional expectations and satisfy (A2.4).

(A2.4) For the $\bar{p}_i$ defined in (A2.3), $\bar{p}_i g_{a0} + g_{ai} = g_{di}$, $i = 1, 2$.

(A2.5) There are continuous and bounded real valued functions $\sigma_{ai}(\cdot)$, $\sigma_{di}(\cdot)$ such that

$$\sigma^{i,\varepsilon}_{a,n+1} = \sigma_{a,i}(X^{\varepsilon}_{S^{i,\varepsilon}_{a,n}}) + \delta_\varepsilon, \qquad \sigma^{i,\varepsilon}_{d,n+1} = \sigma_{d,i}(X^{\varepsilon}_{S^{i,\varepsilon}_{d,n}}) + \delta'_\varepsilon,$$

where δ_ε and $\delta'_\varepsilon \xrightarrow{\varepsilon} 0$ uniformly in all other variables.

In the sequel, we suppose for simplicity that all $\sigma^2_{\alpha,i}(x) > 0$ for all x, α. The results are true even if this condition is violated.

A more convenient representation for $X^\varepsilon(\cdot)$. The second through the fourth terms on the right-hand side of (2.2) go to infinity as $\varepsilon \to 0$. For purposes of the weak convergence analysis, it is helpful to center these terms so that we can work with martingales and processes of bounded variation. We follow closely the procedure used

in [4, § 3] with a slightly different notation. Access to that paper is not needed. Define the following processes:

$$\tilde{A}_0^{i,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{m=1}^{t/\varepsilon} \left(1 - \frac{\alpha_m^i}{\bar{\alpha}_m^i}\right),$$

$$\tilde{A}_0^{0i,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{m=1}^{t/\varepsilon} \left(I_{S_{a,m}^i} - \frac{\bar{p}_i \alpha_m^0}{\bar{\alpha}_m^0}\right), \tag{2.5}$$

$$\tilde{D}_0^{i,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{m=1}^{t/\varepsilon} \left(1 - \frac{\Delta_m^i}{\bar{\Delta}_m^i}\right).$$

The summands in (2.5) are all centered about their conditional expectations with respect to the filtration that measures the "part." Hence, the sums are martingales. Henceforth, we simply write the indicator function that appears in the second sum as $I_m^{i,\varepsilon}$. This is merely for the sake of notational simplicity and is justified by (A2.3).

As in [4, § 3], we can write (recall that $A^{i,\varepsilon}(t) = \bar{S}_a^{i,\varepsilon}(t)/\sqrt{\varepsilon}, \cdots$)

$$A^{i,\varepsilon}(t) = \tilde{A}_0^{i,\varepsilon}(\bar{S}_a^{i,\varepsilon}(t)) + \sqrt{\varepsilon} \sum_{m=1}^{\bar{S}_a^{i,\varepsilon}(t)/\varepsilon} \frac{\alpha_m^i}{\bar{\alpha}_m^i}, \qquad i = 1, 2,$$

$$A^{0i,\varepsilon}(t) = \tilde{A}_0^{0i,\varepsilon}(\bar{S}_a^{0,\varepsilon}(t)) + \sqrt{\varepsilon}\, \bar{p}_i \sum_{m=1}^{\bar{S}_a^{0,\varepsilon}(t)/\varepsilon} \frac{\alpha_m^0}{\bar{\alpha}_m^0}, \tag{2.6}$$

$$D^{i,\varepsilon}(t) = \tilde{D}_0^{i,\varepsilon}(\bar{S}_d^{i,\varepsilon}(t)) + \sqrt{\varepsilon} \sum_{m=1}^{\bar{S}_d^{i,\varepsilon}(t)/\varepsilon} \frac{\Delta_m^i}{\bar{\Delta}_m^i}.$$

The first terms on the right sides of (2.6) are just scaled martingales. The right-hand terms in (2.6) "blow up" as $\varepsilon \to 0$. In (2.2), the sum of the first two minus the third term of (2.6) occurs. Subtracting the far right-hand term on the third line of (2.6) from the sum of far right-hand terms of the first two lines of (2.6), and using the heavy traffic assumption (A2.4), the expansion (A2.1), and the fact that

$$\sqrt{\varepsilon} \sum_{m=1}^{\bar{S}_d^{i,\varepsilon}(t)/\varepsilon} \Delta_m^i = \frac{t}{\sqrt{\varepsilon}} + O(\sqrt{\varepsilon}) = \sqrt{\varepsilon} \sum_{m=1}^{\bar{S}_a^{i,\varepsilon}(t)/\varepsilon} \alpha_m^i,$$

yields (as in [4, § 3]) the expression

$$\varepsilon \sum_{m=1}^{\bar{S}_a^{i,\varepsilon}(t)/\varepsilon} \alpha_m^i a_{im} + \varepsilon \sum_{m=1}^{\bar{S}_a^{0,\varepsilon}(t)/\varepsilon} \bar{p}_i \alpha_m^0 a_{0m} - \varepsilon \sum_{m=1}^{\bar{S}_d^{i,\varepsilon}(t)/\varepsilon} \Delta_m^i d_{im} + \delta^{i,\varepsilon}(t), \tag{2.7}$$

where $\delta^{i,\varepsilon}(\cdot)$ is such that $\sup_{t \le T} |\delta^{i,\varepsilon}(t)| \xrightarrow{\varepsilon} 0$, for each $T < \infty$.

For $i = 1, 2$, let $\tilde{A}^{i,\varepsilon}(t)$, $\tilde{A}^{0i,\varepsilon}(t)$ and $\tilde{D}^{i,\varepsilon}(t)$ denote the first terms on the right sides of (2.6). Define $b_i(x) = a^i(x) + \bar{p}_i a^0(x) - d^i(x)$ and

$$B^{i,\varepsilon}(t) = \int_0^t b_i(X^\varepsilon(s))\, ds.$$

Then, modulo an error (which we absorb into $\delta^{i,\varepsilon}(\cdot)$) of order $O(\varepsilon)$ due to the approximation of the sum by an integral, (2.7) equals $\delta^{i,\varepsilon}(t) + B^{i,\varepsilon}(t)$ and

$$\begin{aligned} X^{i,\varepsilon}(t) = X_0^i + [\tilde{A}^{i,\varepsilon}(t) + \tilde{A}^{i0,\varepsilon}(t) - \tilde{D}^{i,\varepsilon}(t)] \\ + B^{i,\varepsilon}(t) + J^{i,\varepsilon}(t) + Y^{i,\varepsilon}(t) - U^{i,\varepsilon}(t) + \delta^{i,\varepsilon}(t). \end{aligned} \tag{2.8}$$

3. Weak convergence. In this section, we deal with the weak convergence of the terms in (2.8), as $\varepsilon \to 0$. Let $D^k[0, \infty)$ denote the space of R^k-valued right continuous functions with left-hand limits and $C^k[0, \infty)$ the subspace of continuous functions.

For all the weak convergence work, we use $D^k[0,\infty)$ under the Skorokhod topology [5, Chap. 3.5]. We will often use the Skorokhod representation [5, Thm. 3.1.8] so that we can always assume that if a sequence of processes converges weakly, then the convergence is (with probability 1) also pathwise in the topology of the path space.

There are two main problems. First, little is known about the control terms $J^{i,\varepsilon}(\cdot)$. In general, even if bounded, they need not converge in the Skorokhod topology. Indeed, their behavior can be quite "wild." The pseudopath topology [7] could be used, as it has been in [6] for some approximation and convergence questions arising from systems with wide bandwidth noise disturbances under singular controls. For our purposes, it is more convenient to work directly with the Skorokhod topology, but with a rescaled set of processes. (Some comments on the relations between scaling and the pseudopath topology are in [9].) After getting the desired weak convergence, we invert the "limit" of the rescalings to get the result for (2.8).

The second problem concerns the treatment of the reflection terms $Y^{i,\varepsilon}(\cdot)$ and $U^{i,\varepsilon}(\cdot)$. Owing to the presence of the upper boundary, the reflection mapping theorem of [1] and [2] cannot be used directly. The following extension is proved in § 8.

THEOREM 1. *Let Q be a $k\times k$ probability transition matrix whose spectral radius is less than unity. Let $z(\cdot)\in D^k[0,\infty)$ and consider*

$$x(t)=z(t)+(I-Q')y(t)-u(t). \tag{3.1}$$

There is a continuous function (in the topology of uniform convergence on bounded time intervals) $F(\cdot)$ such that $(y(\cdot),u(\cdot))=F(z(\cdot))$ has the following properties: $F(\cdot)$ maps $C^k[0,\infty)$ into $C^k[0,\infty)$ and $D^k[0,\infty)$ into $D^k[0,\infty)$; for $i=1,2$, $y^i(\cdot)$ and $u^i(\cdot)$ are nondecreasing and increase only when $x^i(t)=0$ and $x^i(t)=B_i$, respectively. Equation (3.1) *holds and $x^i(t)\in[0,B_i]$.*

Using the martingale properties of the sums defined in (2.5), it is not hard to prove Theorem 2. In fact, the proof of the first paragraph is given in Lemma 5.2 of [4], and the proof of the second paragraph is in Theorem 5.1 of [4].

THEOREM 2. *Assume* (A2.1)-(A2.5). *Then, for $\alpha=a$ or d, the processes with values $\varepsilon S^{i,\varepsilon}_{a,t/\varepsilon}$ and $\bar S^{i,\varepsilon}_{\alpha}(t)$ converge weakly to the deterministic functions with values $t/g_{\alpha i}$ and $tg_{\alpha i}$, respectively. The processes*

$$\{\tilde A^{1,\varepsilon}(\cdot),\tilde A^{2,\varepsilon}(\cdot),(\tilde A^{01,\varepsilon}(\cdot),\tilde A^{02,\varepsilon}(\cdot)),\tilde D^{1,\varepsilon}(\cdot),\tilde D^{2,\varepsilon}(\cdot),\varepsilon>0\}$$

are tight and the limit of any weakly convergent subsequence of the five sequences (we always pair together $\tilde A^{01,\varepsilon}(\cdot)$ and $\tilde A^{02,\varepsilon}(\cdot)$) are orthogonal continuous martingales.

The quadratic variations of the limit martingales are, respectively, the weak limits of

$$\begin{aligned}&\int_0^t \Sigma_{ai}(X^\varepsilon(s))\,ds, && i=1,2,0\\ &\int_0^t \Sigma_{di}(X^\varepsilon(s))\,ds, && i=1,2,\end{aligned} \tag{3.2}$$

where

$$\Sigma_{ai}(x)=g_{ai}^3\sigma_{ai}^2(x),\qquad i=1,2,$$

$$\Sigma_{a0}(x)=g_{a0}\begin{bmatrix}\bar p_1(1-\bar p_1) & -\bar p_1\bar p_2\\ -\bar p_1\bar p_2 & \bar p_2(1-\bar p_2)\end{bmatrix}+g_{a0}^3\sigma_{a0}^2(x)\begin{bmatrix}1 & \bar p_1\bar p_2\\ \bar p_1\bar p_2 & 1\end{bmatrix},$$

$$\Sigma_{di}(x)=g_{di}^3\sigma_{di}^2(x),\qquad i=1,2.$$

Since the proof of an almost identical result is in the cited reference, we omit it and comment only on how (3.2) is calculated in one case.

The quadratic variation of the discrete parameter martingale $\tilde{A}_0^{i,\varepsilon}(\cdot)$ is (recalling that the argument of $\sigma_{ai}^2(\cdot)$ is the state at the time of arrival of the mth customer)

$$\varepsilon \sum_{m=1}^{t/\varepsilon} E_{a,m}^{i,\varepsilon}\left(1-\frac{\alpha_m^{i,\varepsilon}}{\bar{\alpha}_m^{i,\varepsilon}}\right)^2 = \varepsilon \sum_{m=1}^{t/\varepsilon} g_{ai}^2 \sigma_{ai}^2(X^{\varepsilon}_{S_{a,m}^{i,\varepsilon}}) + (\text{small terms}).$$

Neglecting the small terms (which go to zero, as $\varepsilon \to 0$), we can write the quadratic variation of $\tilde{A}^{i,\varepsilon}(\cdot)$ as

$$\varepsilon \sum_{m=1}^{\bar{S}_a^{i,\varepsilon}(t)/\varepsilon} \frac{g_{ai}^2 \sigma_{ai}^2(X^{\varepsilon}_{S_{a,m}^{i,\varepsilon}})}{\bar{\alpha}_{a,m}^{i,\varepsilon}} \alpha_{a,m}^{i,\varepsilon} + \varepsilon \sum_{m=1}^{\bar{S}_a^{i,\varepsilon}(t)/\varepsilon} \frac{g_{ai}^2 \sigma_{ai}^2(X^{\varepsilon}_{S_{a,m}^{i,\varepsilon}})}{\bar{\alpha}_{a,m}^{i,\varepsilon}} (\bar{\alpha}_{a,m}^{i,\varepsilon} - \alpha_{a,m}^{i,\varepsilon}). \tag{3.3}$$

The variance of the second term in (3.3) is $O(\varepsilon t)$ due to the centering of the summands about the conditional expectations. The first term in (3.3) can be written as (modulo an error of order $O(\sqrt{\varepsilon})$)

$$\int_0^t g_{ai}^3 \sigma_{ai}^2(X^{\varepsilon}(s))\, ds. \tag{3.4}$$

Thus, we obtain the first line of (3.2), for $i = 1, 2$.

The time rescaling. The weak convergence proofs for the terms in (2.8) are facilitated by means of a rescaling or "stretching out" of time. Define $T^{\varepsilon}(\cdot)$ by

$$T^{\varepsilon}(n\varepsilon) = n\varepsilon + \sqrt{\varepsilon} \sum_{m=1}^{n} [\rho_m^{12} + \rho_m^{21} - \rho_m^{12}\rho_m^{21}],$$

and for $t \in (n\varepsilon, n\varepsilon + \varepsilon)$, define $T^{\varepsilon}(t)$ to be the piecewise linear interpolation. Let $\hat{T}^{\varepsilon}(\cdot)$ denote the inverse function to $T^{\varepsilon}(\cdot)$. For any function $\phi(\cdot)$ on $[0, \infty)$, define the function $\hat{\phi}^{\varepsilon}(\cdot)$ by $\hat{\phi}^{\varepsilon}(t) = \phi(\hat{T}^{\varepsilon}(t))$. Similarly, define $\hat{A}^{\alpha,\varepsilon}(t) = \tilde{A}^{\alpha,\varepsilon}(\hat{T}^{\varepsilon}(t))$, etc.

THEOREM 3. *Assume* (A2.1)-(A2.5). *Then*

$$\{\hat{T}^{\varepsilon}(\cdot), \hat{X}^{\varepsilon}(\cdot), \hat{B}^{\varepsilon}(\cdot), \hat{Y}^{i,\varepsilon}(\cdot), \hat{U}^{i,\varepsilon}(\cdot), \hat{J}^{12,\varepsilon}(\cdot), \hat{J}^{21,\varepsilon}(\cdot), \varepsilon > 0\} \tag{3.5}$$

is tight and all limits are continuous processes. Also

$$\{\hat{A}^{1,\varepsilon}(\cdot), \hat{A}^{2,\varepsilon}(\cdot), (\hat{A}^{01,\varepsilon}(\cdot), \hat{A}^{02,\varepsilon}(\cdot)), \hat{D}^{1,\varepsilon}(\cdot), \hat{D}^{2,\varepsilon}(\cdot), \varepsilon > 0\} \tag{3.6}$$

is tight and the limits of any weakly convergent subsequence of the set of five sequences are orthogonal continuous martingales. Let ε index a weakly convergent subsequence of (3.5), (3.6), *and denote the limits by the same letters, but with the ε dropped. Then*

$$\hat{X}^i(t) = X^i(0) + \hat{B}^i(t) + [\hat{A}^i(t) + \hat{A}^{0i}(t) - \hat{D}^i(t)] + \hat{Y}^i(t) - \hat{U}^i(t) + \hat{J}^{ji}(t) - \hat{J}^{ij}(t). \tag{3.7}$$

$\hat{Y}^i(\cdot)$ increases only when $\hat{X}^i(t) = 0$ and $\hat{U}^i(\cdot)$ increases only when $\hat{X}^i(t) = B_i$. Also,

$$\hat{B}^i(t) = \int_0^t b_i(\hat{X}(s))\, d\hat{T}(s). \tag{3.8}$$

The quadratic variations of the martingales are

$$\begin{aligned} &\int_0^t \Sigma_{ai}(\hat{X}(s))\, d\hat{T}(s), \qquad i = 0, 1, 2, \\ &\int_0^t \Sigma_{di}(\hat{X}(s))\, d\hat{T}(s), \qquad i = 1, 2. \end{aligned} \tag{3.9}$$

For the particular chosen weakly convergent subsequence, let $\hat{\mathscr{F}}_t$ denote the minimal σ-algebra that measures $\{\hat{P}(s), s \leqq t\}$, where

$$\hat{P}(s) = (\hat{X}(s), \hat{J}^{12}(s), \hat{J}^{21}(s), \hat{A}^{i}(s), \hat{A}^{0i}(s), \hat{D}^{i}(s), \hat{T}(s), i = 1, 2).$$

Then the martingales are all $\hat{\mathscr{F}}_t$-martingales.

Proof. The set (3.6) is tight and has the asserted properties by Theorem 2, since (3.6) is just the sequence dealt with in Theorem 2, but with a "stretched out" timescale. The $\{\hat{T}^{\varepsilon}(\cdot), J^{12,\varepsilon}(\cdot), J^{21,\varepsilon}(\cdot), \varepsilon > 0\}$ are tight since their increments between any t, $t+s$ are bounded by $s + O(\sqrt{\varepsilon})$. The set $\{\hat{B}^{\varepsilon}(\cdot), \varepsilon > 0\}$ is obviously tight.

To treat the $\hat{Y}^{i,\varepsilon}(\cdot)$ and $\hat{U}^{i,\varepsilon}(\cdot)$, we use the representation of the reflecting terms of Theorem 1. Thus, there is a continuous function (in the sense of Theorem 1) $F_0(\cdot)$ such that

$$(\hat{Y}^{\varepsilon}(\cdot), \hat{U}^{\varepsilon}(\cdot)) = F_0(X_0^{\varepsilon}, \hat{A}^{i,\varepsilon}(\cdot), \hat{A}^{0i,\varepsilon}(\cdot), \hat{D}^{i,\varepsilon}(\cdot), \hat{B}^{i,\varepsilon}(\cdot), \hat{J}^{12,\varepsilon}(\cdot), i = 1, 2).$$

The tightness of $\{\hat{Y}^{\varepsilon}, \hat{U}^{\varepsilon}(\cdot), \hat{X}^{\varepsilon}(\cdot), \varepsilon > 0\}$ and the continuity of the weak limits follows from this and the fact that the argument processes of $F_0(\cdot)$ are tight and have continuous weak limits. Also, the properties asserted below (3.7) hold. The representation (3.8) follows from the equality

$$\hat{B}^{i,\varepsilon}(t) = \int_0^{\hat{T}^{\varepsilon}(t)} b_i(X^{\varepsilon}(s))\, ds = \int_0^t b_i(X^{\varepsilon}(\hat{T}^{\varepsilon}(s)))\, d\hat{T}^{\varepsilon}(s), \tag{3.10}$$

as we will now see. Abusing notation, let ε index a weakly convergent subsequence of the sets in (3.5), (3.6), and suppose that the Skorokhod representation is used so that we can assume that all weak convergences are convergences with probability 1 and are uniform on each bounded time interval (since the limit processes are continuous with probability 1). Since the $\hat{T}^{\varepsilon}(\cdot)$ satisfy $|\hat{T}^{\varepsilon}(t+s) - \hat{T}^{\varepsilon}(t)| = O(s)$, the uniform convergence (on each $[0, t]$) of $\hat{T}^{\varepsilon}(\cdot)$ to continuous $\hat{T}(\cdot)$ and $\hat{X}^{\varepsilon}(\cdot)$ to continuous $\hat{X}(\cdot)$ and (3.10) yield the assertion. A similar proof yields the analogous assertion for the quadratic variation terms.

The last sentence of the theorem is proved in the same way that (5.4) in [4] is proved, via use of the "martingale method," and we only do one case. Let $h(\cdot)$ be an arbitrary real-valued, bounded, and continuous function of its arguments and for arbitrary n, let $t_i \leqq t \leqq t+s$, $i \leqq n$. Define $\hat{P}^{\varepsilon}(t) = (\hat{X}^{\varepsilon}(t), \hat{J}^{12,\varepsilon}(t), \hat{J}^{21,\varepsilon}(t), \hat{A}^{i,\varepsilon}(t), \hat{A}^{0i,\varepsilon}(t), \hat{D}^{i,\varepsilon}(t), \hat{T}^{\varepsilon}(t), i = 1, 2)$. Let ε index a weakly convergent subsequence of $\{\hat{P}^{\varepsilon}(\cdot), \varepsilon > 0\}$. It can be shown that

$$Eh(\hat{P}^{\varepsilon}(t_i), i \leqq n)[\hat{A}^{i,\varepsilon}(t+s) - \hat{A}^{i,\varepsilon}(t)] = 0.$$

This last expression can be shown either by the ideas leading to (5.4) in [4], or by a direct calculation using the definition of the conditional expectation $E_{a,n}^{i,\varepsilon}$ and the fact that the summands in $\hat{A}^{i,\varepsilon}(\cdot)$ are centered about their conditional expectations, given the "past." By the weak convergence and the fact that $\sup_{\varepsilon} E[\hat{A}^{i,\varepsilon}(t)]^2 < \infty$ for each $t < \infty$, we have

$$Eh(\hat{P}(t_i), i \leqq n)[\hat{A}^{i}(t+s) - \hat{A}^{i}(t)] = 0.$$

The arbitrariness of $h(\cdot)$, t_i, n, t, $t+s$, implies that

$$E[\hat{A}(t+s) - \hat{A}(t) \,|\, \hat{P}(u), u \leqq t] = 0,$$

which yields the assertion. □

The inversion of $\hat{T}(\cdot)$. Next we deal with the inversion of the time rescaling $\hat{T}(\cdot)$ to get the appropriate "limits" of the original sets of processes in (2.8). Whether or not this "inversion" can be done depends on the controls. Clearly, if all arrivals at P_0 are rerouted, then for each $t>0$, $T^\varepsilon(t)\to\infty$ as $\varepsilon\to 0$ and $\hat{T}(t)\equiv 0$, and no inversion is possible. However, since the costs associated with this policy go to infinity as $\varepsilon\to 0$, such cases can be excluded. It will turn out that for the controls of practical interest, the inversion can be done.

LEMMA 4. *Assume* (A2.1)-(A2.5) *and that*

$$\sup_\varepsilon \sqrt{\varepsilon}\, E \sum_{m=1}^{t/\varepsilon} [\rho_m^{12}+\rho_m^{21}]<\infty \tag{3.11}$$

for each $t<\infty$. *Then* $\hat{T}(t)<\infty$ *with probability* 1 *for each* $t<\infty$ *and* $\hat{T}(t)\to\infty$ *with probability* 1, *as* $t\to\infty$.

The proof is easy and is omitted.

For each $t>0$, define the random variable

$$T(t)=\min\{\tau\colon \hat{T}(\tau)=t\}.$$

The set $\{T(s), s<\infty\}$ are $\hat{\mathcal{F}}_t$-stopping times, since $\{T(t)\leqq\tau\}=\{\hat{T}(\tau)\geqq t\}\in\hat{\mathcal{F}}_\tau$ for all τ. Define the σ-algebras $\mathcal{F}_t=\hat{\mathcal{F}}_{T(t)}$. For any process $\hat{\phi}(\cdot)$, define the rescaled process $\phi(\cdot)$ by $\phi(t)=\hat{\phi}(T(t))$, except let $\tilde{A}^\alpha(\cdot)$ and $\tilde{D}^\alpha(\cdot)$ denote $\hat{A}^\alpha(T(\cdot))$ and $\hat{D}^\alpha(T(\cdot))$, respectively. Then $\mathcal{F}_t$ is the minimal σ-algebra induced by $\{P(s), s\leqq t\}=\{\hat{P}(T(s)), s\leqq t\}$. The process $T(\cdot)$ is left continuous. Because of this, the $X^i(\cdot)$, $J^{ij}(\cdot)$ in (3.12) will be left continuous. But without loss of generality we can simply take these functions to be right continuous if we wish.

THEOREM 5. *Assume* (A2.1)-(A2.5) *and* (3.11). *Then*

$$X^i(t)=X^i(0)+B^i(t)+[\tilde{A}^i(t)+\tilde{A}^{0i}(t)-\tilde{D}^i(t)]+Y^i(t)-U^i(t)+J^{ji}(t)-J^{ij}(t). \tag{3.12}$$

The $Y^i(\cdot)$ *and* $U^i(\cdot)$ *increase only when* $X^i(t)=0$ ($X^i(t)=B_i$, *respectively*). *The martingales are all* $\mathcal{F}_t$-*martingales. The quadratic variations are given by* (3.9) *with* $\hat{T}(t)$ *replaced by* t *and* $\hat{X}(\cdot)$ *by* $X(\cdot)$.

The proof is just a consequence of Theorem 3, Lemma 4, and the properties of the $T(t)$. The details are omitted.

Remarks on the representation of the martingales. Since the five processes $\tilde{A}^i(\cdot)$, $\tilde{D}^i(\cdot)$, $i=1,2$, and $(\tilde{A}^{01}(\cdot), \tilde{A}^{02}(\cdot))$ are mutually orthogonal martingales, we can represent them as stochastic integrals with respect to mutually independent Wiener processes $w_{ai}(\cdot)$, $w_{di}(\cdot)$. If the $\sigma_{\alpha\beta}$ are never zero (which we have assumed for convenience in this paper), then the $w_{\alpha i}(\cdot)$ are all $\mathcal{F}_t$-Wiener processes. Otherwise, we need to augment the probability space and filtration by adding Wiener processes that are independent of all processes originally defined on the probability space. We can write the martingales in the form

$$\begin{aligned}
&\tilde{A}^i(t)=g_{ai}^{3/2}\int_0^t \sigma_{ai}(X(s))\,dw_{ai}(s)=\int_0^t \Sigma_{ai}^{1/2}(X(s))\,dw_{ai}(s), \qquad i=1,2,\\
&\tilde{D}^i(t)=g_{di}^{3/2}\int_0^t \sigma_{di}(X(s))\,dw_{di}(s)=\int_0^t \Sigma_{di}^{1/2}(X(s))\,dw_{di}(s),\\
&\begin{pmatrix}A^{01}(t)\\ A^{02}(t)\end{pmatrix}=\int_0^t \Sigma_{a0}^{1/2}(X(s))\,dw_{a0}(s).
\end{aligned} \tag{3.13}$$

If the $\{J^{12,\varepsilon}(\cdot), J^{21,\varepsilon}(\cdot), \varepsilon > 0\}$ is tight, then the time change $t \to \hat{T}^{\varepsilon}(t)$ is not needed, and we can work directly with the *original processes* $X^{\varepsilon}(\cdot), \cdots$. We will next give a result for this case that will be useful below. First, we define some new processes by a normalization of the summands in the expressions $\tilde{A}_0^{i,\varepsilon}(\bar{S}_a^{i,\varepsilon}(t))$, $\tilde{A}_0^{0i,\varepsilon}(\bar{S}_a^{0,\varepsilon}(t))$, and $\tilde{D}_0^{i,\varepsilon}(\bar{S}_d^{i,\varepsilon}(t))$ appearing in (2.6). These new processes will actually converge weakly to the Wiener processes $w_{\alpha\beta}(\cdot)$. Define

$$
\begin{aligned}
W_{ai}^{\varepsilon}(t) &= \sqrt{\varepsilon} \sum_{m=1}^{\bar{S}_a^{i,\varepsilon}(t)/\varepsilon} [\Sigma_{ai}(X^{\varepsilon}_{S^i_{a,m}})]^{-1/2} \left(1 - \frac{\alpha^i_m}{\bar{\alpha}^i_m}\right), \qquad i = 1, 2, \\
W_{di}^{\varepsilon}(t) &= \sqrt{\varepsilon} \sum_{m=1}^{\bar{S}_d^{i,\varepsilon}(t)/\varepsilon} [\Sigma_{di}(X^{\varepsilon}_{S^i_{d,m}})]^{-1/2} \left(1 - \frac{\Delta^i_m}{\bar{\Delta}^i_m}\right), \qquad i = 1, 2, \\
W_{a0}^{\varepsilon}(t) &= \sqrt{\varepsilon} \sum_{m=1}^{\bar{S}_a^{0,\varepsilon}(t)/\varepsilon} [\Sigma_{a0}(X^{\varepsilon}_{S^0_{a,m}})]^{-1/2} \begin{Bmatrix} I^{1,\varepsilon}_m - \bar{p}_1 \alpha^0_m / \bar{\alpha}^0_m \\ I^{2,\varepsilon}_m - \bar{p}_2 \alpha^0_m / \bar{\alpha}^0_m \end{Bmatrix}.
\end{aligned} \tag{3.14}
$$

THEOREM 6. *Assume* (A2.1)-(A2.5) *and suppose that* $\{J^{12,\varepsilon}(\cdot), J^{21,\varepsilon}(\cdot), \varepsilon > 0\}$ *is tight. Then* (*note that we pair the two components of* $W_{a0}^{\varepsilon}(\cdot)$)

$$\{X^{\varepsilon}(\cdot), Y^{\varepsilon}(\cdot), U^{\varepsilon}(\cdot), \tilde{A}^{i,\varepsilon}(\cdot), \tilde{A}^{0i,\varepsilon}(\cdot), \tilde{D}^{i,\varepsilon}(\cdot), W_{ai}^{\varepsilon}(\cdot), W_{di}^{\varepsilon}(\cdot), \text{ for all } i\}$$

is tight. Let ε *index a weakly convergent subsequence and denote the limits by the same letters, but without the* ε *superscript. Let* $\mathcal{F}_t$ *be the minimal* σ*-algebra that measures the limit process for* $s \leqq t$. *Then the* $W_{\alpha}(\cdot)$ *are mutually independent standard* $\mathcal{F}_t$*-Wiener processes, and*

$$
\begin{aligned}
\tilde{A}^i(t) &= \int_0^t \Sigma_{ai}^{1/2}(X(s))\, dW_{ai}(s), \\
\tilde{A}^0(t) &= \int_0^t \Sigma_{a0}^{1/2}(X(s))\, dW_{a0}(s), \\
\tilde{D}^i(t) &= \int_0^t \Sigma_{d0}^{1/2}(X(s))\, dW_{d0}(s).
\end{aligned} \tag{3.15}
$$

Also (3.12) *holds.*

Proof. It is easy to show the "Wiener process" result, owing to the centering of the summands and the normalization by the inverse square root of the covariance. The rest is as for Theorem 3, except for the representation (3.15). This can be obtained by using the tightness and a discrete-time approximation, and the details are omitted. □

Service time depending on the customer class. Suppose that the service time distribution depends on the customer type, as it might in Example 3 of § 2. Then, when a customer who was a priori scheduled to P_j is rerouted to $P_i (j \neq i)$, we need to account for the fact that the service time of that customer at P_i might not satisfy (A2.1) and (A2.5). Some minor adjustments are needed in (2.2) and (3.12).

Let $i \neq j$ and let $I_m^{ji,\varepsilon}$ denote the indicator of the event that the mth customer served at P_i was rerouted from P_j. Suppose that there are constants $g_{d,ji}$ such that

$$E[\Delta_m^{i,\varepsilon} \,|\, \text{"past data,"}\ I_m^{ji,\varepsilon} = 1] = g_{d,ji} + O(\sqrt{\varepsilon}).$$

If $g_{d,ji} = g_{di}$ for all i, then the results of this paper hold as stated. In general, it can be shown that neither the $O(\sqrt{\varepsilon})$ above nor the variance of the $\Delta_m^{i,\varepsilon}$ for the rerouted customers appears in the limit equations. This is due to the fact that the fraction of customers that are rerouted goes to zero as $\varepsilon \to 0$, because of the cost of rerouting. The

main problem is with the right-hand term of the third line of (2.6), and we need to correct that term.

We have

$$\sqrt{\varepsilon}\sum_{m=1}^{t/\varepsilon}\frac{\Delta_m^i}{\bar{\Delta}_m^i}=\sqrt{\varepsilon}\sum_{m=1}^{t/\varepsilon}\Delta_m^i[g_{di}+\sqrt{\varepsilon}\,d^i(X^{\varepsilon}_{S^{i,\varepsilon}_{d,m}})+O(\sqrt{\varepsilon})](1-I_m^{ji})$$
$$+\sqrt{\varepsilon}\sum_{m=1}^{t/\varepsilon}\Delta_m^i[g_{d,ji}+O(\sqrt{\varepsilon})]I_m^{ji}.$$

The correction term for the third line of (2.6) is

$$\sqrt{\varepsilon}\sum_{m=1}^{\bar{S}^{i,\varepsilon}_d(t)/\varepsilon}\Delta_m^i[g_{d,ji}-g_{di}]I_m^{ji}=[g_{d,ji}-g_{di}]J^{ji,\varepsilon}(t).$$

All the results of the paper continue to hold if $[g_{d,ji}-g_{di}]J^{ji}(t)$ is subtracted from the right side of (3.12).

4. Boundedness and approximation to $V^{\varepsilon}(x)$. First, we show that there is a control for which the costs are uniformly bounded.

THEOREM 7. *Assume* (A2.1)-(A2.5), *and let* $V^{\varepsilon}(x,0)$ *denote the cost when* $J^{12,\varepsilon}(t)=J^{21,\varepsilon}(t)\equiv 0$. *Then*

$$\sup_{\varepsilon,x}E_xV^{\varepsilon}(x,0)<\infty.$$

Proof. It is enough to prove that

$$\sup_{\varepsilon,n,i}E[U^{i,\varepsilon}(n+1)-U^{i,\varepsilon}(n)]<\infty.$$

Define $M^{i,\varepsilon}(t)=[\tilde{A}^{i,\varepsilon}(t)+\tilde{A}^{0i,\varepsilon}(t)-\tilde{D}^{i,\varepsilon}(t)]$. We let $i=1$, since the proof is the same for $i=2$. For an integer n, define the stopping times (omitting the n and ε-dependence in the notation)

$$\tau_1=\min\{t\geqq n\colon X^{1,\varepsilon}(t)=B_1\},$$
$$\tau_{2m}=\min\{t>\tau_{2m-1}\colon X^{1,\varepsilon}(t)\leqq B_1/2\}\wedge(n+1),$$
$$\tau_{2m+1}=\min\{t>\tau_{2m}\colon X^{1,\varepsilon}(t)=B_1\}\wedge(n+1).$$

Define $N_n^{\varepsilon}=\min\{m\colon \tau_{2m}=n+1\}$. Recall that $U^{1,\varepsilon}(\cdot)$ can increase only on the intervals $[\tau_{2m-1},\tau_{2m}]$ and not on (τ_{2m},τ_{2m+1}). Then

$$U^{1,\varepsilon}(n+1)-U^{1,\varepsilon}(n)=\sum_{m=1}^{N_n^{\varepsilon}+1}[(X^{1,\varepsilon}(\tau_{2m})-X^{1,\varepsilon}(\tau_{2m-1}))$$
$$-(M^{1,\varepsilon}(\tau_{2m})-M^{1,\varepsilon}(\tau_{2m-1}))-(B^{1,\varepsilon}(\tau_{2m})-B^{1,\varepsilon}(\tau_{2m-1}))]. \tag{4.1}$$

By (4.1) and the square integrable martingale property of $M^{i,\varepsilon}(\cdot)$ and the Lipschitz continuity property of $B^{i,\varepsilon}(\cdot)$, there is a constant K_0 such that

$$E|U^{1,\varepsilon}(n+1)-U^{1,\varepsilon}(n)|\leqq K_0+E(N_n^{\varepsilon}+1)B_1. \tag{4.2}$$

Thus, to prove the theorem, we only need bound EN_n^{ε}, uniformly in n and ε.

Given $\alpha_0>0$, there is $\delta_0>0$ such that for all bounded stopping times and for small ε

$$P\left\{\sup_{\tau+\delta_0\geqq s\geqq\tau}[B^{1,\varepsilon}(s)+M^{1,\varepsilon}(s)-(B^{1,\varepsilon}(\tau)+M^{1,\varepsilon}(\tau))]\geqq\frac{B_1}{2}\,\middle|\,\text{data up to }\tau\right\}$$
$$\leqq 1-\alpha_0. \tag{4.3}$$

This implies that

$$P\{\tau_{2m}-\tau_{2m-1}\geqq\delta_0|\text{data up to time }\tau_{2m-1}\}\geqq\alpha_0. \tag{4.4}$$

Consider the problem of a sequence of "Bernoulli" trials, where the conditional probability of success, given the past data, is greater than or equal to α_0 and on each success "time" advances by δ_0. An upper bound for our EN_n^ε is just the mean number of trials that are needed to have $1/\delta_0=n_1$ (the next largest integer) successes. Since the mean number of required trials is monotonic in the (conditional) probability of success, we get an upper bound by assuming that (4.4) is an equality. Then

$$P\{k\text{ trials needed}\}=\binom{k}{n_1}(1-\alpha_0)^{k-n_1}\alpha_0^{n_1},$$

which implies that *all* moments of N_n^ε are bounded, uniformly in n and ε. □

We remark that the proof and the uniform square integrability of the increments in $M^\varepsilon(\cdot)$ and $B^\varepsilon(\cdot)$ (on unit intervals) implies that

$$\sup_{\varepsilon,n}E|U^{i,\varepsilon}(n+1)-U^{i,\varepsilon}(n)|^2<\infty. \tag{4.5}$$

The following corollary will be useful later. It is just a consequence of Theorem 7, the structure of the cost and the discounting. Define $V^\varepsilon(x)=\inf_{J^\varepsilon}V^\varepsilon(x,J^\varepsilon)$.

COROLLARY 8. *Assume* (A2.1)-(A2.5). *Given* $\delta>0$, *there are* $T_0>0$ *and a family of* δ*-optimal controls* $J_\delta^\varepsilon(\cdot)$ *such that* $J_\delta^\varepsilon(\cdot)$ *do not change after time* T_0 (*i.e., after* T_0, *there is no rerouting*).

A very similar proof to that of Theorem 7 yields the following.

THEOREM 9. *Assume* (A2.1)-(A2.5). *If, for each* $t<\infty$

$$\sup_{\varepsilon,T}E[J^{ij,\varepsilon}(t+T)-J^{ij,\varepsilon}(T)]<\infty,\qquad i\neq j,\quad i=1,2, \tag{4.6}$$

then

$$\sup_{\varepsilon,T}E[U^{i,\varepsilon}(t+T)-U^{i,\varepsilon}(T)]<\infty.$$

If

$$\{J^{ij,\varepsilon}(t+T)-J^{ij,\varepsilon}(T),\ \varepsilon>0,\ T<\infty\},\qquad i\neq j,\quad i=1,2, \tag{4.7}$$

is uniformly integrable for each t, *then so is* $\{U^{i,\varepsilon}(t+T)-U^{i,\varepsilon}(T),\ \varepsilon>0,\ T<\infty\}$, $i=1,2$.

5. The limit control problem.

DEFINITION. $J(\cdot)=(J^{12}(\cdot),J^{21}(\cdot))$ is said to be an *admissible control* for the limit controlled reflected diffusion (3.12) if it is nonanticipative with respect to the set of Wiener processes $W(\cdot)=(w_{ai}(\cdot),w_{di}(\cdot),i=1,2,w_{a0}(\cdot))$ that "drive" the martingales $(\tilde{A}^i(\cdot),\tilde{A}^{0i}(\cdot),\tilde{D}^i(\cdot),i=1,2)$ (see the representation (3.13)), and satisfies $J^\alpha(0)=0$, and $J^\alpha(\cdot)$ is nondecreasing, for $\alpha=12$ or 21. We often say simply that the pair $(J(\cdot),W(\cdot))$ is admissible. The cost functional for the limit problem is

$$\begin{aligned}V(x,J,W)=E_x\int_0^\infty e^{-\beta t}k(X(t))\,dt+E_x\int_0^\infty e^{-\beta t}[k_1\,dJ^{12}(t)\\+k_2\,dJ^{21}(t)+c_1\,dU^1(t)+c_2\,dU^2(t)].\end{aligned} \tag{5.1}$$

The $W(\cdot)$ appears in $V(\cdot)$ as well as $J(\cdot)$, since the value of the cost function will depend on the *joint distribution* of $(J(\cdot), W(\cdot))$. Assumption (5.2) in Theorem 10 implies (3.11), and is used to get an inequality for the limit of the costs. Under (A6.2) and without (5.2), the inequality is shown to be an equality.

THEOREM 10. *Assume* (A2.1)-(A2.5) *and that for each n*

$$\sup_{\varepsilon,n,i} E[(J^{12,\varepsilon}(n+1)-J^{12,\varepsilon}(n))+(J^{21,\varepsilon}(n+1)-J^{21,\varepsilon}(n))]<\infty. \tag{5.2}$$

Let ε index a weakly convergent subsequence of (3.5), (3.6) *with limit denoted by* $(\hat{T}(\cdot), \cdots)$. *Let the retransformed processes defined above and in Theorem* 5 *be denoted by* $(T(\cdot), \cdots)$. *Then*

$$\underline{\lim}_{\varepsilon} V^{\varepsilon}(x, J^{\varepsilon}) \geqq V(x, J, W), \tag{5.3}$$

where $W(\cdot)=(w_{ai}(\cdot), w_{di}(\cdot), i=1,2)$ *is the Wiener process that is used to represent the martingales* (*see* (3.13)). *If*

$$\{J^{12,\varepsilon}(n+1)-J^{12,\varepsilon}(n), J^{21,\varepsilon}(n+1)-J^{21,\varepsilon}(n), \varepsilon>0, n<\infty\} \tag{5.4}$$

is uniformly integrable, then

$$V^{\varepsilon}(x, J^{\varepsilon}) \to V(x, J, W). \tag{5.5}$$

Proof. The hypothesis (5.2) implies that $\inf_{\varepsilon} E\hat{T}^{\varepsilon}(t)\to\infty$ and $E\hat{T}(t)\to\infty$ as $t\to\infty$. Thus, the "inverse" transformation $T(\cdot)$ is well defined. It also implies that we need only work on a finite interval (see Corollary 8). For simplicity, we work with only a couple of the terms of the cost functional. We have

$$\begin{aligned} &\int_0^{\infty} e^{-\beta t}k(X^{\varepsilon}(t))\,dt=\int_0^{\infty} e^{-\beta\hat{T}^{\varepsilon}(t)}k(\hat{X}^{\varepsilon}(t))\,d\hat{T}^{\varepsilon}(t),\\ &\int_0^{\infty} e^{-\beta t}\,dJ^{12,\varepsilon}(t)=\int_0^{\infty} e^{-\beta\hat{T}^{\varepsilon}(t)}\,d\hat{J}^{12,\varepsilon}(t). \end{aligned} \tag{5.6}$$

By the weak convergence and the argument of Theorem 3, the right sides of (5.6) converge in distribution to the left sides of

$$\begin{aligned} &\int_0^{\infty} e^{-\beta\hat{T}(t)}k(\hat{X}(t))\,d\hat{T}(t)=\int_0^{\infty} e^{-\beta t}k(X(t))\,dt,\\ &\int_0^{\infty} e^{-\beta\hat{T}(t)}\,d\hat{J}^{12}(t)=\int_0^{\infty} e^{-\beta t}\,dJ^{12}(t). \end{aligned} \tag{5.7}$$

The left sides of (5.7) equal the right sides of (5.7) by the rescaling. The theorem follows from the cited convergences (together with those for the other components of the cost) and Fatous' lemma. □

Theorems 3 and 5 imply that every limit of a weakly convergent subsequence is a legitimate control problem in the sense that the pair $(J(\cdot), W(\cdot))$ that occurs in the representation of the limit is admissible. This fact and Theorem 10 imply the following theorem.

THEOREM 11. *Assume* (A2.1)-(A2.5). *Let* $J^{12,\varepsilon}(\cdot)$, $J^{21,\varepsilon}(\cdot)$ *denote the optimal controls for the physical process. Define*

$$V^{\varepsilon}(x)=\inf_{J^{\varepsilon}} V^{\varepsilon}(x, J^{\varepsilon}), \qquad V(x)=\inf_{(J,W)\text{adm}} V(x, J, W).$$

Then

$$\underline{\lim}_{\varepsilon} V^{\varepsilon}(x) \geqq V(x). \tag{5.8}$$

Remark. We note that (5.2) can be assumed in Theorem 11. If it does not hold for the optimal policy, for each $\delta > 0$ it will hold for the *δ-optimal policy*, owing to the discounting and Corollary 8. We want to prove that (5.8) is an equality. To get the equality, we will need to use the fact that $V^\varepsilon(x)$ is actually an *optimal* cost. To do this, first we need to study approximations to the control problem for the limit model (3.12), (5.1). We will show that there is an optimal policy for the limit, and that it can be approximated by a policy that we can apply to the $X^\varepsilon(\cdot)$ process, and that will be "recovered" under the weak convergence. Such results will get us the desired equality in (5.8) (Theorem 17), together with a basis for an effective computational approximation. The computational methods and associated proofs will be dealt with in a subsequent paper.

6. Approximations for the limit problem, and convergence of the costs. To prove equality in (5.8), we first establish the existence of an optimal policy for (3.12), (5.1), and then obtain a sequence of approximations to the optimal control. We will use the following assumption.

(A6.1) $k(\cdot)$, $b_i(\cdot)$, $\sigma_{ia}(\cdot)$, $\sigma_{id}(\cdot)$ are continuous.

THEOREM 12. *Assume* (A6.1). *Consider the limit control problem* (3.12), (5.1). *There is an optimal policy $\bar{J}(\cdot)$ in the sense that there is $(\bar{X}(\cdot), \bar{J}(\cdot), \bar{W}(\cdot), \cdots)$ satisfying* (3.12), *where $\bar{W}(\cdot) = \{\bar{w}_{ai}(\cdot), \bar{w}_{di}(\cdot), i = 1, 2\}$ "drives" the martingales $\tilde{A}^i(\cdot), \cdots$, as in* (3.13) *and the pair $(\bar{J}(\cdot), \bar{W}(\cdot))$ is admissible and*

$$V(x, \bar{J}, \bar{W}) \leqq V(x, J, W),$$

for all admissible pairs $(J(\cdot), W(\cdot))$.

Proof. The proof is very similar to those of Theorems 3 and 5, and we make only a few comments. Let $(J^n(\cdot), W^n(\cdot))$ be an admissible pair for (3.12), and write the corresponding form of (3.12) as $(j \neq i)$

$$\begin{aligned} X^{i,n}(t) = X_0^i + B^{i,n}(t) + [\tilde{A}^{i,n}(t) + \tilde{A}^{0i,n}(t) - \tilde{D}^{i,n}(t)] \\ + [J^{ij,n}(t) - J^{ji,n}(t)] + Y^{i,n}(t) - U^{i,n}(t)], \end{aligned} \tag{6.1}$$

where $B^{i,n}(t) = \int_0^t b_i(X^{i,n}(s))\, ds$ and $W^n(\cdot) = \{w_{ai}^n(\cdot), w_{di}^n(\cdot), i = 1, 2\}$ "drives" the martingales $\tilde{A}^{i,n}(\cdot), \cdots$, as in (3.13). Let $(J^n(\cdot), W^n(\cdot))$ be a minimizing sequence in that $V(x, J^n, W^n) \downarrow V(x)$.

By Theorems 7 and 10, $V(x) < \infty$. Hence

$$\sup_n E[J^{12,n}(t) + J^{21,n}(t)] < \infty,$$

for each $t < \infty$. Define the time change

$$T^n(t) = t + J^{12,n}(t) + J^{21,n}(t),$$

and the inverse $\hat{T}^n(t) = \min\{\tau\colon T^n(\tau) = t\}$. Analogous to the notation used in Theorem 3, define $\hat{X}^n(\cdot) = X^n(\hat{T}^n(\cdot)), \cdots$. Then

$$\begin{aligned} \hat{X}^{i,n}(t) = X_0^i + \hat{B}^{i,n}(t) + [\hat{A}^{i,n}(t) + \hat{A}^{i0,n}(t) - \hat{D}^{i,n}(t)] \\ + \hat{J}^{ij,n}(t) - \hat{J}^{ji,n}(t) + \hat{Y}^{i,n}(t) - \hat{U}^{i,n}(t). \end{aligned}$$

As in Theorem 3, there is a function $F_0(\cdot)$ that maps $C^k[0, \infty)$ into $C^k[0, \infty)$, for the appropriate integer k and is continuous in the topology of uniform convergence on bounded time intervals and is such that for all n

$$(\hat{Y}^n(\cdot), \hat{U}^n(\cdot)) = F_0(X_0, \hat{A}^{i,n}(\cdot), \hat{A}^{i0,n}(\cdot), \hat{D}^{i,n}(\cdot), \hat{B}^{i,n}(\cdot), \hat{J}^n(\cdot), i = 1, 2).$$

The set $\{\hat{T}^n(\cdot), \hat{X}^n(\cdot), \hat{J}^n(\cdot), \hat{A}^{i,n}(\cdot), \hat{A}^{i0,n}(\cdot), \hat{D}^{i,n}(\cdot), \hat{B}^{i,n}(\cdot), i=1,2, n<\infty\}$ is tight. Abusing notation, let n index a weakly convergent subsequence with limit denoted by $(\hat{T}(\cdot), \hat{X}(\cdot), \cdots)$. As in Theorem 3 the $(\hat{A}^1(\cdot), \hat{A}^2(\cdot), (\hat{A}^{10}(\cdot), \hat{A}^{20}(\cdot)), \hat{D}^1(\cdot), \hat{D}^2(\cdot))$ are orthogonal continuous martingales with quadratic variation defined by (3.9). Define the inverse scaling $T(t)=\min\{\tau: \hat{T}(\tau)=t\}$, and the rescaled processes $X(t)=\hat{X}(T(t)), \cdots$. Then (3.12) holds, and the martingales have the representation (3.13) with respect to some Wiener process $W(\cdot)=(w_{ai}(\cdot), w_{di}(\cdot), i=1,2)$ such that the pair $(J(\cdot), W(\cdot))$ is admissible. By an argument that is almost identical to that of Theorem 10, we have

$$\underline{\lim}_n V(x, J^n, W) \geqq V(x, J, W). \tag{6.2}$$

We must have the equality in (6.2) since $V(x, J^n, W^n)\downarrow V(x)$. Thus, $(J(\cdot), W(\cdot))$ is an optimal admissible pair. □

The following lemma, whose proof is similar to that of Theorems 7 and 9, will be useful later.

THEOREM 13. *Assume* (A6.1), *and let* $(J_n(\cdot), W_n(\cdot))$ *be admissible, with* $X_n(\cdot)$, $Y_n(\cdot)$, *and* $U_n(\cdot)$, *the associated state and reflection process. If*

$$\{J_n^{ij}(t+T)-J_n^{ji}(T), n<\infty\}, \qquad j\neq i, \quad j=1,2,$$

is uniformly integrable, then so is

$$\{U_n^i(t+T)-U_n^i(T), n<\infty\}, \qquad j\neq i, \quad j=1,2.$$

THEOREM 14. *Assume* (A6.1) *and for small* $\delta>0$ *let* $(J_0(\cdot), W_0(\cdot))$ *be a* δ*-optimal admissible pair, with* $X_0(\cdot)$ *being the associated solution to* (3.12). *Define* $\tau_N=\sup\{t: J_0^{12}(t)\leqq N, J_0^{21}(t)\leqq N\}$, *and let* $J^N(\cdot)$ *be the policy that equals* $J_0(\cdot)$ *until* τ_N, *and is constant thereafter. Write the solution to* (3.12) *as*

$$\begin{aligned} X^{i,N}(t) = X^i(0) + B^{i,N}(t) + [\tilde{A}^{i,N}(t) + \tilde{A}^{0i,N}(t) - \tilde{D}^{i,N}(t)] \\ + Y^{i,N}(t) - U^{i,N}(t) + J^{ji,N}(t) - J^{ij,N}(t). \end{aligned}$$

Let $W_0^N(\cdot)$ *be the set of Wiener processes that "drives" the martingales* $(\tilde{A}^{i,N}(\cdot), \cdots)$. *Then, as* $N\to\infty$,

$$V(x, J_0^N, W_0^N)\to V(x, J_0, W_0). \tag{6.3}$$

Proof. We can suppose without loss of generality that there is a $T<\infty$ such that $J_0(\cdot)$ is constant after T (by an argument similar to that leading to Corollary 8). Since $J^{i,N}(T)\uparrow J_0^i(T)$, and $EJ_0^i(T)<\infty$, the $\{J_0^N(T), N<\infty\}$ is uniformly integrable, and so is $\{U^{i,N}(n+1)-U^{i,N}(n), n<\infty, N<\infty, i=1,2\}$ by Theorem 13. Since $\tau_N\uparrow\infty$, we can suppose that

$$(X_0^N(\cdot), J_0^N(\cdot), W_0^N(\cdot))\to(X_0(\cdot), J_0(\cdot), W_0(\cdot))$$

pathwise. The theorem follows from this convergence, the cited uniform integrability, and an argument similar to that in Theorem 10. □

DEFINITION. A solution $X(\cdot)$ to (3.12), (3.13) is said to be *unique in the weak sense* if the distribution of the admissible pair $(J(\cdot), W(\cdot))$ determines that of $(J(\cdot), W(\cdot), X(\cdot))$.

To obtain our approximation results we require that for each $\delta>0$, there is a δ-optimal control that gives a well-defined solution of (3.12), (3.13).

(A6.2) For each $\delta>0$, there is a δ-optimal control $J(\cdot)$ for which $(J(\cdot), W(\cdot))$ is admissible for some $W(\cdot)$ and the corresponding solution $X(\cdot)$ to (3.12), (3.13) is unique in the weak sense.

In the next theorem, we show that there is a δ-optimal control that is bounded, piecewise constant, and jumps "in increments."

THEOREM 15. *Assume* (A6.1)-(A6.2) *and let* $(J(\cdot), W(\cdot))$ *be a* δ*-optimal pair satisfying* (A6.2), *with* $X(\cdot)$ *denoting the corresponding solution process. For* $\Delta > 0$ *and* $\rho > 0$, *define the control* $J_{\Delta\rho}(\cdot)$ *as the piecewise constant control satisfying* $dJ^{ij}_{\Delta\rho}(t) = 0$ *on the interval* $(n\Delta, n\Delta+\Delta)$ *and on* $[0, \Delta)$. *For* $k \geqq 0$ *and* $n \geqq 1$, *set* $dJ^{ij}_{\Delta\rho}(n\Delta) = k\rho$ *if* $J^{ij}(n\Delta) - J^{ij}(n\Delta-\Delta) \in [k\rho, k\rho+\rho)$. *Then*

$$\lim_{\Delta,\rho} V(x, J_{\Delta\rho}, W) = V(x, J, W). \tag{6.4}$$

Proof. By Theorem 13, we can suppose that $J(\cdot)$ is uniformly bounded. By construction, $(J_{\Delta\rho}(\cdot), W(\cdot))$ is an admissible pair. A solution to (3.12), (3.13) can be constructed on some probability space for some pair having the same distribution as $(J_{\Delta\rho}(\cdot), W(\cdot))$, and we use the same notation for that new probability space. Let $(X_{\Delta\rho}(\cdot), U_{\Delta\rho}(\cdot), Y_{\Delta\rho}(\cdot))$ denote the corresponding state and reflection processes. The set

$$\{X_{\Delta\rho}(\cdot), J_{\Delta\rho}(\cdot), W(\cdot), U_{\Delta\rho}(\cdot), Y_{\Delta\rho}(\cdot), \Delta > 0, \rho > 0\} \tag{6.5}$$

is tight and the weak limits all satisfy (3.12), (3.13). By the uniqueness (A6.2), the limit of any weakly convergent subsequence of the set (6.5) satisfies (3.12), (3.13). Then (6.4) follows from the weak convergence and the boundedness of $J(\cdot)$ and the consequent uniform integrability of $\{U^i_{\Delta\rho}(n+1) - U^i_{\Delta\rho}(n),\ n < \infty,\ \Delta > 0,\ \rho > 0,\ i = 1, 2\}$. □

For $\Delta > 0$, $\rho > 0$, let $\mathcal{T}_{\Delta\rho}$ denote the set of admissible (with respect to some given Wiener process $W(\cdot)$) controls that are bounded, are constant on each interval $[n\Delta, n\Delta+\Delta)$, jump only at the times $n\Delta$, and where $J^{ij}(n\Delta) - J^{ij}(n\Delta^-)$ is an integral multiple of ρ. By Theorem 15 and (A6.2), we know that for each $\delta > 0$ there are $\Delta > 0$, $\rho > 0$ such that there is a δ-optimal control in some $\mathcal{T}_{\Delta\rho}$. We will need to define this control in such a way that it can be used for the physical $X^\epsilon(\cdot)$ process.

Write $k = (k_1, k_2)$, a multi-index, where k_i is either an integer or zero. Fix the Wiener process $W(\cdot)$ and Δ and ρ. For $J(\cdot) \in \mathcal{T}_{\Delta\rho}$, $(J(\cdot), W(\cdot))$ is admissible. For $\gamma > 0$ and integers k and n, define $q_{nk\gamma}(\cdot)$ by

$$\begin{aligned} q_{nk\gamma}(J(m\Delta), m < n, W(l\gamma), l\gamma \leqq n\Delta) &= P\{dJ(n\Delta) \\ &= k\rho \,|\, J(m\Delta), m < n, W(l\gamma), l\gamma \leqq n\Delta\}. \end{aligned} \tag{6.6}$$

By the martingale convergence theorem, as $\gamma \to 0$

$$q_{nk\gamma}(J(m\Delta), m < n, W(l\gamma), l\gamma \leqq n\Delta) \to P\{dJ(n\Delta) = k\rho \,|\, J(m\Delta), m < n, W(s), s \leqq n\Delta\}$$

with probability 1 (Wiener measure) for each k, n and value of $\{J(m\Delta), m < n\}$.

For each $\gamma > 0$, we next choose a control $J_\gamma(\cdot) \in \mathcal{T}_{\Delta\rho}$ recursively by means of the following set of conditional probabilities:

$$\begin{aligned} P\{dJ_\gamma(n\Delta) = k\rho \,|\, J_\gamma(m\Delta), m < n, W(s), s \leqq n\Delta\} \\ = q_{nk\gamma}(J_\gamma(m\Delta), m < n, W(l\gamma), l\gamma \leqq n\Delta). \end{aligned} \tag{6.7}$$

Equation (6.7) specifies the joint law of admissible $(J_\gamma(\cdot), W(\cdot))$, and there is a solution $X(\cdot)$ on some sample space that is associated with a pair with the same distribution as $(J_\gamma(\cdot), W(\cdot))$.

We will need the following condition.

(A6.3) The uncontrolled system ($J(t) \equiv 0$) has a unique (in the weak sense) solution for each initial condition.

THEOREM 16. *Assume* (A6.1)-(A6.3). *Let* $(J(\cdot), W(\cdot))$ *be admissible with* $J(\cdot) \in \mathcal{T}_{\Delta\rho}$ *for some* $\Delta > 0$, $\rho > 0$. *Define* $J_\gamma(\cdot)$ *as above. Then*

$$V(x, J_\gamma, W) \to V(x, J, W). \tag{6.8}$$

The function $q_{nk\gamma}(J_\gamma(m\Delta), m < n, \cdot)$, *which is used to get* $J_\gamma(\cdot)$, *can be chosen to be continuous for each* n, k, γ, *and values of the set* $\{J_\gamma(m\Delta), m < n\}$.

Proof. The proof of (6.8) follows from the weak convergence $\{J_\gamma(\cdot), W(\cdot), \gamma > 0\} \Rightarrow (J(\cdot), W(\cdot))$, as $\gamma \to 0$ and the uniform boundedness of the controls. By (A6.3), the solution to (3.12), (3.13) is unique in the weak sense for any admissible $J(\cdot)$ in $\mathcal{T}_{\Delta\rho}$. The last sentence of the theorem follows from the fact that for each n, k, γ, and value of $\{J_\gamma(m\Delta), m < \Delta\}$, we can approximate $q_{nk\gamma}(J_\gamma(m\Delta), m < n, \cdot)$ by a sequence of continuous distribution functions that converge to $q_{nk\gamma}(J_\gamma(m\Delta), m < n, \cdot)$ with probability 1 (Wiener measure). □

The optimality theorem. We now return to the physical process (2.8), and prove equality in (5.8). Let $(J_\gamma(\cdot), W(\cdot))$ be admissible with $J_\gamma(\cdot)$ chosen by (6.7), where the $q_{nk\gamma}(\cdot)$ have the continuity property asserted in Theorem 15. Recall the definition of $W^\varepsilon(\cdot)$ given above Theorem 6.

We would like to define a control $J^\varepsilon(\cdot)$ for $X^\varepsilon(\cdot)$ such that $\{J^\varepsilon(\cdot), W^\varepsilon(\cdot)\}$ converges weakly to $(J_\gamma(\cdot), W(\cdot))$ as $\varepsilon \to 0$. First, consider the control $\tilde{J}^\varepsilon(\cdot)$ defined as follows, where the $q_{nk\gamma}(\cdot)$ are continuous in the W-arguments; $\tilde{J}^\varepsilon(\cdot)$ is constant on the intervals $[n\Delta, n\Delta + \Delta)$, and

$$\begin{aligned} &P\{d\tilde{J}^\varepsilon(n\Delta) = k\rho \,|\, \tilde{J}^\varepsilon(m\Delta), m < n, W^\varepsilon(s), s \leqq n\Delta\} \\ &\quad = P\{d\tilde{J}^\varepsilon(n\Delta) = k\rho \,|\, \tilde{J}^\varepsilon(m\Delta), m < n, W^\varepsilon(i\gamma), i\gamma \leqq n\Delta\} \\ &\quad = q_{nk\gamma}(\tilde{J}^\varepsilon(m\Delta), m < n, W^\varepsilon(i\gamma), i\gamma \leqq n\Delta). \end{aligned} \tag{6.9}$$

The control law (6.9) cannot quite be realized for $X^\varepsilon(\cdot)$, since the controls for $X^\varepsilon(\cdot)$ are the result of rerouting decisions and $X^\varepsilon(\cdot)$ *cannot* be impulsively controlled. But we can come close enough to realizing the above $\tilde{J}^\varepsilon(\cdot)$, as follows.

For notational simplicity, let the $k_i\rho$, $i = 1, 2$, be integral multiples of $\sqrt{\varepsilon}$. Let $\Delta_\varepsilon \to 0$ as $\varepsilon \to 0$ such that $\Delta_\varepsilon/\sqrt{\varepsilon} \to \infty$. Let Q_ε^n denote the event that there are at least equal to $(B_1 + B_2)/\sqrt{\varepsilon}$ arrivals at P_0 on $[n\Delta, n\Delta + \Delta_\varepsilon)$. We have $P\{Q_\varepsilon^n\} \to 1$ as $\varepsilon \to 0$. Define $J^\varepsilon(\cdot)$ to be any control with the following properties: $J^\varepsilon(\cdot)$ is constant on $[n\Delta + \Delta_\varepsilon, n\Delta + \Delta)$ and on $[0, \Delta)$; for $n > 0$, the rerouting on $[n\Delta, n\Delta + \Delta_\varepsilon)$ is such that

$$\begin{aligned} &P\{J^\varepsilon(n\Delta + \Delta_\varepsilon) - J^\varepsilon(n\Delta) = k\rho \,|\, J^\varepsilon(m\Delta), m < n, W^\varepsilon(s), s \leqq n\Delta, Q_\varepsilon^n\} \\ &\quad = P\{J^\varepsilon(n\Delta + \Delta_\varepsilon) - J^\varepsilon(n\Delta) = k\rho \,|\, J^\varepsilon(m\Delta), m < n, W^\varepsilon(i\gamma), i\gamma \leqq n\Delta, Q_\varepsilon^n\} \\ &\quad = q_{nk\gamma}(J^\varepsilon(m\Delta), m < n, W^\varepsilon(i\gamma), i\gamma \leqq n\Delta). \end{aligned} \tag{6.10}$$

The limit of the costs associated with this just constructed $J^\varepsilon(\cdot)$ is the same as if the jumps of $J^\varepsilon(\cdot)$ are at the time $n\Delta + \Delta_\varepsilon$, $n = 1, 2, \cdots$, only and not spreadout over $[n\Delta, n\Delta + \Delta_\varepsilon)$. This can be easily proved by the "time charge method." Under this "new" $J^\varepsilon(\cdot)$, $\{J^\varepsilon(\cdot), W^\varepsilon(\cdot)\}$ clearly converges weakly to $(J_\gamma(\cdot), W(\cdot))$ as $\varepsilon \to 0$. Now, the above discussion and Theorems 6, 10, and 11 yield the following theorem, where we use the $J^\varepsilon(\cdot)$ just described. Note that the rescalings are not necessary, due to the fixed form of $J^\varepsilon(\cdot)$; we get weak convergence directly in the Skorokhod topology.

THEOREM 17. *Assume* (A2.1)-(A2.5) *and* (A6.1)-(A6.3). *Then* $\{X^\varepsilon(\cdot), Y^\varepsilon(\cdot), U^\varepsilon(\cdot), W^\varepsilon(\cdot), J^\varepsilon(\cdot)\}$ *converges weakly to* $(X(\cdot), Y(\cdot), U(\cdot), W(\cdot), J(\cdot))$, *where* $W(\cdot) = (w_{ai}(\cdot), w_{di}(\cdot), i = 0, 1, 2)$ *and the limit satisfies* (3.12), (3.13). *Also*

$$V^\varepsilon(x, J^\varepsilon) \to V(x, J_\gamma, W), \tag{6.11}$$

$$V^\varepsilon(x) \to V(x). \tag{6.12}$$

Remarks. In [4], it was shown that certain forms of nearly optimal controls for the limit process were also nearly optimal for the physical process under heavy traffic. A similar situation holds here, and this partly justifies the use of the heavy traffic limit, but we reserve the comments for a future paper on numerical methods.

Suppose that $k_1 > 0$ but $k_2 < 0$ and $|k_2| < k_1$. Then a very similar analysis can be carried out with similar results. The costs $V^\varepsilon(x)$ can be bounded from below since the profit to be made by rerouting from P_1 to P_2 is bounded, due to the limited idle time at P_2.

7. A more general network model. The general controlled routing open network version of Fig. 1 can be treated for any number of servers. Because of the notational burden involved in writing all the possible "rerouting terms," we give the extension only to the model of Fig. 2, which differs from Fig. 1 only in that feedback is allowed. We continue to use the notation of the previous sections, except for the following additions. Let $I_n^{ij,\varepsilon}$ be the indicator of the event that a service completed at P_i $(i \neq 0)$ at real time n is routed to P_j $(j \neq i)$ if $j \neq 0$, and leaves the network if $j = 0$. The input from P_j to P_i is

$$D^{ji,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{m=1}^{t/\varepsilon} \psi_m^j I_m^{ji}, \qquad j = 1, 2, \quad i = 0, 1, 2.$$

$D^{j0,\varepsilon}(t)$ denotes the scaled number of outputs of P_j that leave the system directly. The "fictitious" outputs from P_j (which are due to our convention of P_j "processing" even with an empty queue) that are sent to P_i are

$$Y^{ji,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{m=1}^{t/\varepsilon} \psi_m^j I_m^{ji} I_{\{X_m^{j,\varepsilon}=0\}}.$$

The overflow due to a full buffer at P_i is

$$U^{i,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{m=1}^{t/\varepsilon} \xi_m^i + \sqrt{\varepsilon} \sum_{j \neq i} \sum_{m=1}^{t/\varepsilon} [\xi_m^0 (I_m^i + I_m^j \rho_m^{ji} - I_m^i \rho_n^{ij}) + \psi_m^j I_m^{ji}] I_{\{X_m^{i,\varepsilon} = B_i\}}.$$

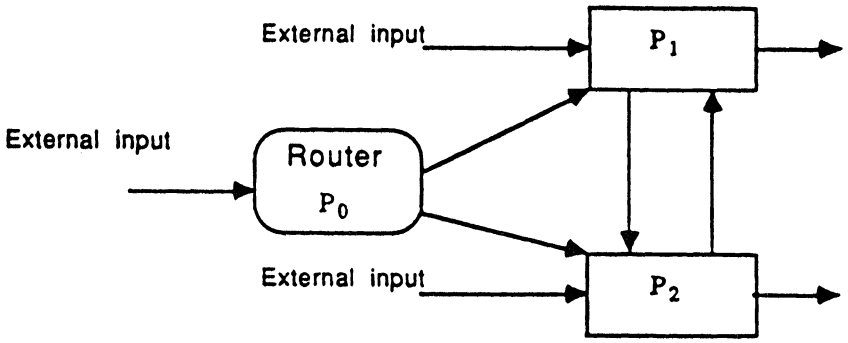


FIG. 2. *A routing problem with feedback.*

Then, for $j \neq i$, $j \neq 0$,

$$X^{i,\varepsilon}(t) = X_0^{i,\varepsilon} + A^{i,\varepsilon}(t) + A^{i0,\varepsilon}(t) - D^{i0,\varepsilon}(t) - D^{ij,\varepsilon}(t) + D^{ji,\varepsilon}(t) + Y^{i,\varepsilon}(t) - Y^{ji,\varepsilon}(t) - U^{i,\varepsilon}(t) + J^{ji,\varepsilon}(t) - J^{ij,\varepsilon}(t). \tag{7.1}$$

Note that it is possible for a customer to be lost during processing, e.g., if a customer from P_i is sent to $P_j, j \neq i$, when the buffer of P_j is full. If this is not desirable, then it can be handled in several ways. We can separate out such losses from the others and put a high cost on them. An alternative is to block P_i by holding the customer until there is room at P_j. If an "impulsive" cost is associated with this blocking, then the limit would be a singularly *and* impulsively controlled system. The model could be altered so that "external arrivals" are not admitted to P_j if the buffer is more than $(B_i - \Delta)/\sqrt{\varepsilon}$ full, for some $\Delta > 0$. The weak convergence techniques and the general results will be quite similar to those in this paper.

We continue to use the cost functional (2.3).

Replace (A2.3) and (A2.4) by the following:

(A2.3′) There are $\bar{p}_{ij} < 1$ such that ($\bar{p}_{0i}$ replaces the $\bar{p}_i$ of (A2.3))

$$P\{I_n^{ij,\varepsilon} = 1 \,|\, \text{all arrival and service intervals and routings starting by time } n, \text{ except for } I_n^{ij,\varepsilon}\} = \bar{p}_{ij}.$$

(A2.4′) $g_{di} = g_{ai} + \bar{p}_{0i} g_{a0} + \bar{p}_{ji} g_{dj}$, $j \neq i$, $i, j \neq 0$.

In analogy to the definitions of the centered processes $\tilde{A}_0^{i,\varepsilon}(\cdot)$, $\tilde{D}_0^{i,\varepsilon}(\cdot)$ given in (2.5), define ($j \neq 0$)

$$\tilde{D}_0^{ji,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{m=1}^{t/\varepsilon} [I_{S_{d,m}^i}^{ji} - \bar{p}_{ji} \Delta_m^j / \bar{\Delta}_m^j].$$

Define the centered reflection process

$$\tilde{Y}^{ji,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{m=1}^{t/\varepsilon} \psi_m^j (I_m^{ji} - \bar{p}_{ji}) I_{\{X_m^{j,\varepsilon} = 0\}}.$$

Define $\tilde{D}^{ij}(t) = \tilde{D}_0^{ij,\varepsilon}(\bar{S}_d^{i,\varepsilon}(t))$. By the same method that was used to get (2.8), we can write (7.1) in the form ($i \neq 0$, $j \neq 0$, $j \neq i$)

$$\begin{aligned} X^{i,\varepsilon}(t) = {} & X^{i,\varepsilon}(0) + B^{i,\varepsilon}(t) + \tilde{A}^{i,\varepsilon}(t) + \tilde{A}^{i0,\varepsilon}(t) - \tilde{D}^{i0,\varepsilon}(t) - \tilde{D}^{ij,\varepsilon}(t) \\ & + \tilde{D}^{ji,\varepsilon}(t) + Y^{i,\varepsilon}(t) - \bar{p}_{ji} Y^{j,\varepsilon}(t) - \tilde{Y}^{ji,\varepsilon}(t) - U^{i,\varepsilon}(t) \\ & + [J^{ji,\varepsilon}(t) - J^{ij,\varepsilon}(t)] + \delta^{i,\varepsilon}(t), \end{aligned} \tag{7.2}$$

where $\delta^{i,\varepsilon}(\cdot)$ is as in (2.7).

THEOREM 18. *Assume* (A2.1), (A2.2), (A2.3′), (A2.4′), (A2.5). *Then the five sets of processes* (*we pair* (A^{01}, A^{02}), (D^{10}, D^{12}), (D^{20}, D^{21}))

$$\{\tilde{A}^{1,\varepsilon}(\cdot), \tilde{A}^{2,\varepsilon}(\cdot), (\tilde{A}^{01,\varepsilon}(\cdot), \tilde{A}^{02,\varepsilon}(\cdot)), (\tilde{D}^{10,\varepsilon}(\cdot), \tilde{D}^{12,\varepsilon}(\cdot)), (\tilde{D}^{20,\varepsilon}(\cdot), \tilde{D}^{21,\varepsilon}(\cdot)), \varepsilon > 0\} \tag{7.3}$$

are tight. The limits are continuous martingales. Theorems and Lemmas 2–16 *hold, with the obvious changes necessitated by the additional terms in* (7.3). *The limit reflected diffusion is*

$$\begin{aligned} X^i(t) = {} & X^i(0) + \tilde{A}^i(t) + \tilde{A}^{0i}(t) - \tilde{D}^{i0}(t) - \tilde{D}^{ij}(t) + \tilde{D}^{ji}(t) - B^i(t) \\ & + J^i(t) - U^i(t) + Y^i(t) - \bar{p}_{ji} Y^i(t), \qquad i \neq j. \end{aligned} \tag{7.4}$$

Also, for $i \neq j$

$$\text{quad var}\begin{pmatrix}\tilde{D}^{i0}(t)\\ \tilde{D}^{ij}(t)\end{pmatrix} = \int_0^t \Sigma_{dij}(X(s))\,ds,$$

where

$$\Sigma_{dij}(x) = g_{di}\begin{bmatrix}\bar{p}_{i0}(1-\bar{p}_{i0}) & -\bar{p}_{i0}\bar{p}_{ij}\\ -\bar{p}_{i0}\bar{p}_{ij} & \bar{p}_{ij}(1-\bar{p}_{ij})\end{bmatrix} + g_{di}^3\sigma_{di}^2(x)\begin{bmatrix}1 & \bar{p}_{i0}\bar{p}_{ij}\\ \bar{p}_{i0}\bar{p}_{ij} & 1\end{bmatrix}. \tag{7.5}$$

Proof. All the details are copies of what was done in Theorems 2-16 and Lemmas 2-16, except for the treatment of the $\tilde{Y}^{ij,\varepsilon}(\cdot)$ term, and some details in Theorem 7. Define $\hat{Y}^{ij,\varepsilon}(t) = \tilde{Y}^{ij,\varepsilon}(\hat{T}^{\varepsilon}(t))$. The summands in the $\hat{Y}^{ij,\varepsilon}(\cdot)$ are centered about their conditional expectations, given the "past," and hence $\hat{Y}^{ij,\varepsilon}(\cdot)$ is a martingale sum. Its variance is bounded by

$$\varepsilon E \sum_{m=0}^{\hat{T}^{\varepsilon}(t)/\varepsilon} I_{\{X_n^{i,\varepsilon}=0\}} = O(t). \tag{7.6}$$

Since the summands (without the $\sqrt{\varepsilon}$ included) in $\hat{Y}^{ij,\varepsilon}(\cdot)$ are uniformly square integrable, $\{\hat{Y}^{ij,\varepsilon}(\cdot), \varepsilon > 0\}$ is tight, and all weak limits are continuous processes. Write the scaled form of (7.2) as

$$\hat{X}^{i,\varepsilon}(t) \equiv X_0^{i,\varepsilon} + \hat{Z}^{i,\varepsilon}(t) + [\hat{Y}^{i,\varepsilon} - \bar{p}_{ji}\hat{Y}^{j,\varepsilon}(t)] - \hat{U}^{i,\varepsilon}(t) \tag{7.7}$$

with the obvious definition of $\hat{Z}^{i,\varepsilon}(\cdot)$. $\{\hat{Z}^{i,\varepsilon}(\cdot), \varepsilon > 0\}$ is tight and all weak limits are continuous processes. Thus, by Theorem 1, $\{\hat{Y}^{i,\varepsilon}(\cdot), \hat{U}^{i,\varepsilon}(\cdot), \varepsilon > 0\}$ is tight and all weak limits are continuous processes. This implies that for each $t < \infty$,

$$\sqrt{\varepsilon} \sum_{m=0}^{\hat{T}^{\varepsilon}(t)/\varepsilon} I_{\{X_m^{i,\varepsilon}=0\}}$$

is bounded in probability uniformly in ε, for otherwise some subsequence of $\hat{Y}^{i,\varepsilon}(t)$ would go to infinity with a positive probability. Hence the left side of (7.6) goes to zero as $\varepsilon \to 0$ for each $t < \infty$. Thus $\hat{Y}^{ij,\varepsilon}(\cdot) \Rightarrow$ zero process.

Theorem 7 also continues to hold, since in the present case we write (4.1) as

$$[U^{1,\varepsilon}(n+1) + Y^{21,\varepsilon}(n+1)] - [U^{1,\varepsilon}(n) + Y^{21,\varepsilon}(n)] = \text{right side of (4.1).}$$

The left-hand side of (4.3) now becomes

$$P\Big\{\sup_{\tau+\delta_0 \geqq s \geqq \tau} [(B^{1,\varepsilon}(s) + M^{1,\varepsilon}(s) - Y^{21,\varepsilon}(s)) - (B^{1,\varepsilon}(\tau) + M^{1,\varepsilon}(\tau) - Y^{21,\varepsilon}(\tau))| \geqq \delta_0] \Big| \text{ data up to } \tau\Big\}.$$

Since $Y^{21,\varepsilon}(\cdot)$ is nondecreasing, the expression is still $\leqq 1 - \alpha_0$ for small enough δ_0, and we can continue the proof of Theorem 7. □

8. Proof of Theorem 1. For notational simplicity, we prove the theorem for $k = 2$ and then comment on the general case. The proof in the general case is the same in all essentials. The following result is proved in [1] and [2].

LEMMA. *Let P be a degenerate Markov transition matrix whose spectral radius is less than unity. Then there is a unique "nonanticipative" function $\tilde{F}(\cdot)$ with the following properties: $\tilde{F}(\cdot)$ maps $D^k[0,\infty)$ into $D^k[0,\infty)$ and $C^k[0,\infty)$ into $C^k[0,\infty)$, and is continuous in the topology of uniform convergence on bounded intervals. Let $\tilde{z}(\cdot)\in D^k[0,\infty)$ and define $\tilde{y}(\cdot)=(\tilde{y}^i(\cdot), i\leqq k)=\tilde{F}(\tilde{z}(\cdot))$. Define $\tilde{x}(t)=\tilde{z}(t)+(I-P')(\tilde{y}(t))$. The $\tilde{y}^i(\cdot)$ are nondecreasing and $\tilde{y}^i(\cdot)$ can increase only when the $\tilde{x}^i(t)=0$. Also $\tilde{x}^i(t)\geqq 0$, for all i, t.*

To prove Theorem 1, we will use the lemma in a "sequential" way. Refer to Fig. 3. We can assume without loss of generality that the diagonal entries in Q in (3.1) are zero. There are four different reflection maps that appear in (3.1), depending on which segment of the boundary is involved. On the boundary $(d, a, b)\equiv$ segment 1, our system (3.1) is

$$x(t)=z(t)+\begin{bmatrix}1 & -q_{21}\\ -q_{12} & 1\end{bmatrix}y(t), \qquad x^i(t)\geqq 0. \tag{8.1}$$

For the system (8.1), with the constraint $x^i(t)\geqq 0$, the lemma defines a continuous mapping $z(\cdot)\to(y(\cdot), u(\cdot))$, where $u(\cdot)\equiv 0$. Call this mapping $F_1(\cdot)$. On the other segments, the system is

$$x(t)=z(t)+\begin{bmatrix}0 & -q_{21}\\ 0 & 1\end{bmatrix}y(t)-\begin{pmatrix}u^1(t)\\ 0\end{pmatrix}, \qquad \text{segment } 2=(a, b, c), \tag{8.2}$$

$$x(t)=z(t)-\begin{pmatrix}u^1(t)\\ u^2(t)\end{pmatrix}, \qquad \text{segment } 3=(b, c, d), \tag{8.3}$$

$$x(t)=z(t)+\begin{bmatrix}1 & 0\\ -q_{12} & 0\end{bmatrix}y(t)-\begin{pmatrix}0\\ u^2(t)\end{pmatrix}, \qquad \text{segment } 4=(c, d, a). \tag{8.4}$$

The reflection maps for (8.2) to (8.4) (with the sides extended to ∞) are trivial, as we will now see, since they are each just concatenations of two one-dimensional applications of the lemma. Let $F_2(\cdot)$ denote the map associated with (8.2) that sends $z(\cdot)$ into $(y(\cdot), u(\cdot))$. $F_2(\cdot)$ is constructed as follows. First, we have $y^1(\cdot)=u^2(\cdot)=0$.

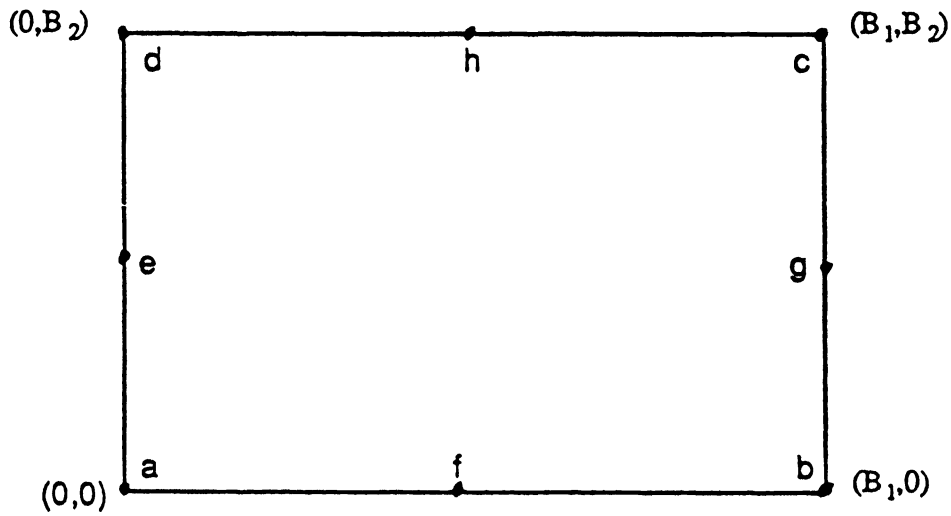


FIG. 3. *Boundary sections for the proof of Theorem* 1.

Then $y^2(\cdot)$ is defined by the lemma for $k=1$; in particular, $y^2(t)=-\min\{0, \inf_{s\leqq t} z^2(s)\}$. Finally, the $u^1(\cdot)$ is defined by the reflection needed to keep $x^1(t)\leqq B_1$; i.e., $0\leqq B_1-x^1(t)$ or

$$u^1(t)=-\min\left\{0, \inf_{s\leqq t}(B_1-z^1(s)+q_{21}y^2(s))\right\}. \tag{8.5}$$

Similarly, we define the analogous continuous map $F_3(\cdot)$ and $F_4(\cdot)$ associated with (8.3) and (8.4), respectively. The calculation of the $y(\cdot)$ and $u(\cdot)$ always decouples into two separate calculations, first getting $y(\cdot)$ and then getting $u(\cdot)$, even for the general k case as seen below.

Define $S=[0, B_1]\times[0, B_2]$. Let $z(0)\in S$ without loss of generality. Define $x(\cdot)$, $y_n(\cdot)$, $u_n(\cdot)$, $z_n(\cdot)$, $\delta y_n(\cdot)$, $\delta u_n(\cdot)$, and τ_n as follows: $\tau_0=0$, $z_0(\cdot)=z(\cdot)$, $\delta u_0(\cdot)=\delta y_0(\cdot)=0$, $\tau_1=\inf\{t: z(t)\notin S\}$, $u(t)=y(t)=0$ on $[0, \tau_1]$ and $x(t)=x_0(t)=z(t)$ on $[0, \tau_1]$. In general, for $n\geqq 1$, given τ_n, $x_{n-1}(\cdot)$ and $y(\cdot)$ and $u(\cdot)$ on $[0, \tau_n]$, define:

$$\begin{aligned}
&z_n(t)=z(t)+[I-Q']y(t\wedge\tau_n)-u(t\wedge\tau_n),\\
&s_n=\text{a boundary segment (1, 2, 3, or 4) on which } x_{n-1}(\tau_n) \text{ lies},\\
&(\delta u_n(\cdot), \delta y_n(\cdot))=F_{s_n}(z_n(\cdot)),\\
&x_n(t)=z_n(t)+[I-Q']\delta y_n(t)-\delta u_n(t),\\
&x(t)=x_n(t) \quad \text{for } t\leqq\tau_{n+1},\\
&\tau_{n+1}=\inf\{t: x_n(t)\notin S\},\\
&u(t)=u(\tau_n)+\delta u_n(t) \quad \text{for } t\in[\tau_n, \tau_{n+1}],\\
&y(t)=y(\tau_n)+\delta y_n(t) \quad \text{for } t\in[\tau_n, \tau_{n+1}].
\end{aligned} \tag{8.6}$$

Note that $z_n(t)\in S$ until at least time τ_n and $x_n(t)\in S$ until at least time τ_{n+1}. Hence $(\delta u_n(t), \delta y_n(t))=0$ until at least τ_n. Also $x_n(t)=z_n(t)$, $t\leqq\tau_n$, $z_{n+1}(t)=z_n(t)$ on $[0, \tau_{n+1}]$, $x_{n+1}(t)=x_n(t)$ on $[0, \tau_{n+1}]$. The idea in constructing $x_n(\cdot)$ is that when $x_n(\cdot)$ exits S on a certain boundary segment, we use the reflection map (F_1, F_2, F_3, or F_4, as appropriate) for that segment to get $x_{n+1}(\cdot)$, until $x_{n+1}(\cdot)$ exits S. It must exit on a "different" segment. $F_{s_n}(\cdot)$ can be any map associated with the boundary segment on which the exit point $x_{n-1}(\tau_n)$ lies. Except for in the corner points, there are two such maps associated with each point on the boundary. Which map is chosen is immaterial. For definiteness, choose $F_1(\cdot)$ on $[e, a, f]$, $F_2(\cdot)$ on $(f, b, g]$, $F_3(\cdot)$ on $(g, c, h]$ and $F_4(\cdot)$ on (h, d, e). We can verify that $\tau_n\to\infty$ and (by induction using the lemma) that the constructed $x(\cdot)$, $y(\cdot)$, $u(\cdot)$ satisfy Theorem 1. To see that the choice of the map used at the points f, g, h, e is immaterial, let $x_{n-1}(\tau_n)=f$. Then $F(z_n(\cdot))=F_2(z_n(\cdot))$ until the infimum of the times that $x_n(\cdot)$ leaves S through $[b, c, d, a]$. An induction argument based on this observation shows that the choices at the points f, g, h, e, are immaterial.

Remark on the general $k>2$ case. There is always a decomposition of the construction of $(\delta y_n(\cdot), \delta u_n(\cdot))$ into the two sequential steps: first calculate $\delta y_n(\cdot)$ via the lemma, for a reduced system; then calculate the $\delta u_n^i(\cdot)$ individually via an appropriate analogue of (8.5). Just to illustrate this point, consider the case where the space is $S=[0, B]^k$, and focus on the faces of S meeting in the corner $(B, B, 0, 0, \cdots, 0)$. On

these faces (excluding the edges that do not touch $(B, B, 0, 0, \cdots, 0)$), (3.1) is

$$x(t) = z(t) + \begin{bmatrix} 0 & 0 & -q_{31} & \cdots & -q_{k1} \\ 0 & 0 & -q_{32} & \cdots & -q_{k2} \\ & & & & \\ 0 & & \tilde{I} - \tilde{Q}' & & \end{bmatrix} y(t) - \begin{pmatrix} u^1(t) \\ u^2(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where $\tilde{I} - \tilde{Q}'$ is a reduced transition matrix. Then first get $(y^2(\cdot), \cdots, y^k(\cdot))$ from the lemma, and then define (for $i = 1, 2$)

$$u^i(t) = -\min\left\{0, \inf_{s \leq t}\left(B - z^i(t) + \sum_{j=3}^{k} q_{ji} y^j(t)\right)\right\}.$$

REFERENCES

[1] M. I. REIMAN, *Open queueing networks in heavy traffic*, Math. Oper. Res., 9 (1984), pp. 441–458.

[2] J. M. HARRISON AND M. I. REIMAN, *Reflected Brownian motion on an orthant*, Ann. Probab., 9 (1981), pp. 302–308.

[3] D. L. IGLEHART AND W. WHITT, *Multiple channel queues in heavy traffic*, Adv. in Appl. Math., 2 (1970), pp. 150–177.

[4] H. J. KUSHNER AND K. M. RAMACHANDRAN, *Optimal and approximately optimal control policies for queues in heavy traffic*, SIAM J. Control Optim., 27 (1989), pp. 1293–1318.

[5] S. N. ETHIER AND T. G. KURTZ, *Markov Processes; Characterization and Convergence*, John Wiley, New York, 1986.

[6] H. J. KUSHNER AND K. M. RAMACHANDRAN, *Nearly optimal singular controls for wideband noise driven systems*, SIAM J. Control Optim., 25 (1987), pp. 289–315.

[7] P. A. MEYER AND W. A. ZHENG, *Tightness criteria for laws of semimartingales*, Ann. Inst. H. Poincaré, 20 (1984), pp. 353–372.

[8] S. E. SHREVE, J. P. LEHOSZKY, AND D. P. GAVER, *Optimal consumption for general diffusion with absorbing and reflecting barriers*, SIAM J. Control Optim., 22 (1984), pp. 55–75.

[9] T. KURTZ, *Meyer-Zheng tightness and almost uniform convergence*, preprint, Mathematics Department, University of Wisconsin, Madison, WI, 1987.

[10] J. M. HARRISON, *Brownian models of queueing networks with heterogeneous customer populations*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, W. H. Fleming and J.-P. Lions, eds., IMA Volumes in Mathematics and Its Applications 10, 1988.

[11] G. J. FOSCHINI, *On heavy traffic diffusion analysis and dynamic routing in packet switched networks*, in Computer Performance, K. M. Chandy and M. Reiser, eds., North-Holland, Amsterdam, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 5, pp. 1234-1250, September 1990

011

A ZERO-SUM DIFFERENTIAL GAME IN A FINITE DURATION WITH SWITCHING STRATEGIES*

JIONGMIN YONG†

Abstract. A zero-sum differential game of finite horizon with both players using switching controls is studied. Positive switching costs are associated with each player. Under some suitable conditions, it is proved that the Elliot-Kalton upper and lower value functions of the game are the unique viscosity solution of the same Isaacs' equation, which turns out to be a system of evolutionary quasi-variational inequalities with bilateral obstacles. The existence of the Elliot-Kalton value of the game then follows. Some limiting cases are also discussed.

Key words. differential games, switching strategies, value, Isaacs equation, quasi-variational inequality, dynamic programming

AMS(MOS) subject classifications. 90D25, 90C39, 49C20

1. Introduction. In this paper, we consider a differential game of the following type:

$$\dot{y}(t) = g(t, y(t), a(t), b(t)), \qquad t \in [0, T], \tag{1.1}$$

where $a(\cdot)$ and $b(\cdot)$ are piecewise constant functions of the following type:

$$a(\cdot) = \sum_{i \geq 1} a_{i-1} \chi_{[\theta_{i-1}, \theta_i)}(\cdot),$$

$$b(\cdot) = \sum_{j \geq 1} b_{j-1} \chi_{[\tau_{j-1}, \tau_j)}(\cdot).$$

In the game, the first player uses control $a(\cdot)$ from some class to minimize the payoff functional

$$J(a(\cdot), b(\cdot)) = \int_0^T f(t, y(t), a(t), b(t))\, dt + \sum_{i \geq 1} k(\theta_i, a_{i-1}, a_i) - \sum_{j \geq 1} l(\tau_j, b_{j-1}, b_j), \tag{1.2}$$

and the second player uses control $b(\cdot)$ to maximize the payoff above. In (1.2), the integral term represents the running cost of the game and the other two summation terms represent the switching costs for the players I and II, respectively. Here, $k(\cdot, \cdot, \cdot)$ and $l(\cdot, \cdot, \cdot)$ are some given nonnegative functions. The main feature of the problem is that at each moment of changing the values of $a(\cdot)$ (respectively, $b(\cdot)$) from one to another, there is a strictly positive cost associated with it. Thus, k and l are called switching cost functions. Due to the appearance of these positive switching costs, our differential game is different from the classical one ([1], [3], [4], [10], [11], [13], [20]). In [22], we studied a similar problem, but it was of infinite horizon and autonomous. Thus, the Isaacs equation was stationary and the uniqueness of the viscosity solution was relatively easy to get. In this paper, however, the problem is of finite horizon and is nonautonomous. Thus, the Isaacs equation will be of evolutionary type. On the other hand, we allow the maps g and f to have certain growth rates (we note here that in [22], these maps were supposed to be bounded). We adopt some techniques provided in [12] and modify the idea we used in [16] and [22] to get the uniqueness of the viscosity solution of the Isaacs equation, which, in turn, gives the existence of the value for our differential game.

* Received by the editor May 17, 1989; accepted for publication (in revised form) November 12, 1989. This work was partially supported by Chinese National Natural Science Foundation grant 0188416.

† Department of Mathematics, Fudan University, Shanghai 200433, China.

2. Preliminaries. As in [22], we let $A=\{1, 2, \cdots, m\}$, $B=\{1, 2, \cdots, n\}$, $T>0$, and X be a finite-dimensional Euclidean space. Let $g:[0, T]\times X\times A\times B\to X$, $f:[0, T]\times X\times A\times B\to\mathbb{R}$, $h:X\to\mathbb{R}$, $k:[0, T]\times A\times A\to\mathbb{R}^+\equiv[0, \infty)$, $l:[0, T]\times B\times B\to\mathbb{R}^+$ be continuous functions satisfying the following hypotheses:

(H1) There exist $L>0$ and a strictly increasing continuous function $\omega(\cdot, \cdot):\mathbb{R}^+\times\mathbb{R}^+\to\mathbb{R}^+$ with $\omega(r, 0)=0$, for all $r\geqq 0$, such that for all $x, \hat{x}\in X$, $t\in[0, T]$, $(a, b)\in A\times B$,

$$|g(t, x, a, b)-g(t, \hat{x}, a, b)|\leqq L|x-\hat{x}|, \tag{2.1}$$

$$|g(t, x, a, b)|\leqq L(1+|x|), \tag{2.2}$$

$$|f(t, x, a, b)-f(t, \hat{x}, a, b)|\leqq\omega(|x|+|\hat{x}|, |x-\hat{x}|), \tag{2.3}$$

$$|f(t, 0, a, b)|\leqq L, \tag{2.4}$$

$$|h(x)-h(\hat{x})|\leqq\omega(|x|+|\hat{x}|, |x-\hat{x}|), \tag{2.5}$$

$$|h(0)|\leqq L. \tag{2.6}$$

(H2) For all $a, \hat{a}, \tilde{a}\in A$, $a\neq\hat{a}\neq\tilde{a}$, and $0\leqq s\leqq t\leqq T$,

$$k(t, a, \tilde{a})<k(t, a, \hat{a})+k(t, \hat{a}, \tilde{a}), \tag{2.7}$$

$$k(t, a, \hat{a})>0, \qquad k(t, a, a)=0, \tag{2.8}$$

$$k(t, a, \tilde{a})\leqq k(s, a, \tilde{a}). \tag{2.9}$$

(H3) For all $b, \hat{b}, \tilde{b}\in B$, $b\neq\hat{b}\neq\tilde{b}$, and $0\leqq s\leqq t\leqq T$,

$$l(t, b, \tilde{b})<l(t, b, \hat{b})+l(t, \hat{b}, \tilde{b}), \tag{2.10}$$

$$l(t, b, \hat{b})>0, \qquad l(t, b, b)=0, \tag{2.11}$$

$$l(t, b, \tilde{b})\leqq l(s, b, \tilde{b}). \tag{2.12}$$

Remark 2.1. The constant L in (H1) can be replaced by some function in $L^1(0, T)$. Also, the moduli of continuity of f and h (i.e., the functions involved in (2.3) and (2.5)) can be different. On the other hand, we take the constant in (2.4) and (2.6) the same as L just for the later notational simplicity. Conditions (2.9) and (2.12) are adopted from [16]. They will play important roles in the proof of the continuity of the lower and upper value functions of the game.

Next, for $s\in[0, T)$, $a\in A$, $b\in B$, we define the following control sets:

$$\mathscr{A}^{a,s}=\Big\{a(\cdot)=\sum_{i\geqq 1}a_{i-1}\chi_{[\theta_{i-1},\theta_i)}(\cdot):[s, T]\to A\,|\,a_0=a,\ \theta_0=s,\ \theta_i\in[s, T],$$
$$\forall i\geqq 1;\ \theta_i\uparrow T,\ a_{i+1}\neq a_i,\ \text{if } \theta_{i+1}<T;\ \sum_{i\geqq 1}k(\theta_i, a_{i-1}, a_i)<\infty\Big\},$$

$$\mathscr{B}^{b,s}=\Big\{b(\cdot)=\sum_{j\geqq 1}b_{j-1}\chi_{[\tau_{j-1},\tau_j)}(\cdot):[s, T]\to B\,|\,b_0=b,\ \tau_0=s,\ \tau_j\in[s, T],$$
$$\forall j\geqq 1;\ \tau_j\uparrow T,\ b_{j+1}\neq b_j,\ \text{if } \tau_{j+1}<T;\ \sum_{j\geqq 1}l(\tau_j, b_{j-1}, b_j)<\infty\Big\}.$$

For any $a(\cdot)\in\mathscr{A}^{a,s}$, from

$$\sum_{i\geqq 1}k(\theta_i, a_{i-1}, a_i)<\infty,$$

we know that there exists an integer $K = K_{a(\cdot)}$, such that

$$a(\cdot) = \sum_{i=1}^{K} a_{i-1}\chi_{[\theta_{i-1},\theta_i)}(\cdot) + a_K\chi_{[\theta_K,T)}(\cdot) \qquad (\theta_K < T).$$

The same thing holds for any $b(\cdot) \in \mathcal{B}^{b,s}$. Hereafter, we keep in mind that $\sum_{i\geq 1}$ and $\sum_{j\geq 1}$ are just finite sums. Also, we assume that there are no switchings made at time T for both players I and II because we understand that the game is terminated as soon as $t = T$.

The following definition is adopted from [8] (see also [9], [22]).

DEFINITION 2.2. For given $s \in [0, T)$, $a \in A$ (respectively, $b \in B$), an admissible strategy $\alpha^{a,s}$ (respectively, $\beta^{b,s}$) for player I (respectively, II) on $[s, T]$ is a mapping $\alpha^{a,s} : \cup_{b\in B} \mathcal{B}^{b,s} \to \mathcal{A}^{a,s}$ (respectively, $\beta^{b,s} : \cup_{a\in A} \mathcal{A}^{a,s} \to \mathcal{B}^{b,s}$), such that

$$b(t) = \hat{b}(t) \quad (\text{resp. } a(t) = \hat{a}(t)) \quad \forall t \in [s, \hat{s}],$$

implies

$$\alpha^{a,s}[b(\cdot)](t) = \alpha^{a,s}[\hat{b}(\cdot)](t) \quad (\text{resp. } \beta^{b,s}[a(\cdot)](t) = \beta^{b,s}[\hat{a}(\cdot)](t)) \quad \forall t \in [s, \hat{s}].$$

We denote all admissible strategies for player I (respectively, II) on $[s, T]$ by $\Gamma^a[s, T]$ (respectively, $\Delta^b[s, T]$). We take the convention that

$$\mathcal{A}^{a,T} = \{a\}, \qquad \Gamma^a[T, T] = \{a\},$$
$$\mathcal{B}^{b,T} = \{b\}, \qquad \Delta^b[T, T] = \{b\}.$$

It is clear that for any $b(\cdot) \in \mathcal{B}^{b,s}$ (respectively, $a(\cdot) \in \mathcal{A}^{a,s}$) and $\alpha \in \Gamma^a[s, T]$ (respectively, $\beta \in \Delta^b[s, T]$), we have

$$\alpha[b(\cdot)] \in \mathcal{A}^{a,s} \ (\text{resp. } \beta[a(\cdot)] \in \mathcal{B}^{b,s}).$$

On the other hand, for any $(a, b) \in A \times B$, $x \in X$, $s \in [0, T)$, and $(a(\cdot), b(\cdot)) \in \mathcal{A}^{a,s} \times \mathcal{B}^{b,s}$, by (H1), we know that there exists a unique solution $y(\cdot)$ of the following problem:

$$\begin{aligned} &\dot{y}(t) = g(t, y(t), a(t), b(t)), \qquad t \in (s, T], \\ &y(s) = x. \end{aligned} \tag{2.13}$$

Here, $y(\cdot)$ depends on s, x, $a(\cdot)$, and $b(\cdot)$. We denote $y_{s,x}(\cdot) \equiv y(\cdot)$ to emphasize the dependence of $y(\cdot)$ on (s, x) and we always keep in mind that $y_{s,x}(\cdot)$ also depends on $a(\cdot)$ and $b(\cdot)$. Then, we consider the following payoff functional:

$$\begin{aligned} J^{a,b}_{s,x}(a(\cdot), b(\cdot)) = &\int_s^T f(t, y_{s,x}(t), a(t), b(t))\, dt + h(y_{s,x}(T)) \\ &+ \sum_{i\geq 1} k(\theta_i, a_{i-1}, a_i) - \sum_{j\geq 1} l(\tau_j, b_{j-1}, b_j). \end{aligned} \tag{2.14}$$

Above and in the following whenever terms such as the right-hand side of (2.14) appear together, we always understand that

$$a(\cdot) = \sum_{i\geq 1} a_{i-1}\chi_{[\theta_{i-1},\theta_i)}(\cdot), \qquad a_0 = a,$$

$$b(\cdot) = \sum_{j\geq 1} b_{j-1}\chi_{[\tau_{j-1},\tau_j)}(\cdot), \qquad b_0 = b,$$

i.e., $\{\theta_i, a_i\}$ and $\{\tau_j, b_j\}$ are associated with $a(\cdot)$ and $b(\cdot)$, respectively. Also, by our convention,

$$J^{a,b}_{T,x}(a(\cdot), b(\cdot)) = h(x) \quad \forall (a, b, x) \in A \times B \times X. \tag{2.15}$$

From the above analysis, we see that for any $(s, x, a, b) \in [0, T] \times X \times A \times B$, $b(\cdot) \in \mathcal{B}^{s,b}$ and $\alpha \in \Gamma^a[s, T]$, we can find $y_{s,x}(\cdot)$, the unique solution of (2.13) corresponding to the control pair $(\alpha[b(\cdot)], b(\cdot))$. Thus, the payoff $J^{a,b}_{s,x}(\alpha[b(\cdot)], b(\cdot))$ is well defined. Then, we define

$$\begin{aligned} &V^{a,b}(s, x) = \inf_{\alpha \in \Gamma^a[s,T]} \sup_{b(\cdot) \in \mathcal{B}^{b,s}} J^{a,b}_{s,x}(\alpha[b(\cdot)], b(\cdot)), \\ &V^{a,b}(T, x) = h(x). \end{aligned} \tag{2.16}$$

Similarly, we define

$$\begin{aligned} &U^{a,b}(s, x) = \sup_{\beta \in \Delta^b[s,T]} \inf_{a(\cdot) \in \mathcal{A}^{a,s}} J^{a,b}_{s,x}(a(\cdot), \beta[a(\cdot)]), \\ &U^{a,b}(T, x) = h(x). \end{aligned} \tag{2.17}$$

We call the $(m \times n)$-matrix-valued functions $V(s, x)$ and $U(s, x)$ lower and upper (Elliot-Kalton) value functions of our differential game, respectively (cf. [8], [9], [22], [23]).

Now, let us give some basic properties of the lower and upper value functions.

LEMMA 2.3. *For any* $(a, b) \in A \times B$, $s, \bar{s} \in [0, T]$, *and* $x, \hat{x} \in X$,

$$|V^{a,b}(s, x)|, |U^{a,b}(s, x)| \leqq \bar{C}(|x|), \tag{2.18}$$

$$|V^{a,b}(s, x) - V^{a,b}(s, \hat{x})|, |U^{a,b}(s, x) - U^{a,b}(s, \hat{x})| \leqq \omega_J(|x| + |\hat{x}|, |x - \hat{x}|), \tag{2.19}$$

$$|V^{a,b}(s, x) - V^{a,b}(\bar{s}, x)|, |U^{a,b}(s, x) - U^{a,b}(\bar{s}, x)| \leqq \bar{C}(|x|)|s - \bar{s}|, \tag{2.20}$$

where $\bar{C}(\cdot): \mathbb{R}^+ \to \mathbb{R}^+$, $\omega_J(\cdot, \cdot): \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ *with* ω_J *being continuous and* $\omega_J(r, 0) = 0$, *for all* $r \geqq 0$.

Proof. For any $(a(\cdot), b(\cdot)) \in \mathcal{A}^{a,s} \times \mathcal{B}^{b,s}$ and $x, \hat{x} \in S$, if we let $y(\cdot)$ and $\hat{y}(\cdot)$ be the solutions of (2.13) corresponding to $(a(\cdot), b(\cdot), x)$ and $(a(\cdot), b(\cdot), \hat{x})$, respectively, then, by the Gronwall inequality, we have

$$\begin{aligned} &|y(t)| \leqq (|x| + 1)\, e^{LT}, \qquad t \in [s, T], \\ &|y(t) - \hat{y}(t)| \leqq e^{LT}|x - \hat{x}|, \qquad t \in [s, T]. \end{aligned}$$

Thus, we have

$$\begin{aligned} &|J^{a,b}_{s,x}(a(\cdot), b(\cdot)) - J^{a,b}_{s,\hat{x}}(a(\cdot), b(\cdot))| \\ &\quad \leqq \int_s^T |f(t, y(t), a(t), b(t)) - f(t, \hat{y}(t), a(t), b(t))|\, dt + |h(y(T)) - h(\hat{y}(T))| \\ &\quad \leqq \int_s^T \omega(|y(t)| + |\hat{y}(t)|, |y(t) - \hat{y}(t)|)\, dt + \omega(|y(T)| + |\hat{y}(T)|, |y(T) - \hat{y}(T)|) \\ &\quad \leqq (1 + T)\omega(e^{LT}(2 + |x| + |\hat{x}|), e^{LT}|x - \hat{x}|). \end{aligned}$$

Thus, (2.19) follows for some ω_J. To obtain (2.18), let us first observe the following:

$$\begin{aligned} I &\equiv \int_s^T |f(t, y(t), a(t), b(t))|\, dt + |h(y(T))| \\ &\leqq \int_s^T [L + \omega(|y(t)|, |y(t)|)]\, dt + L + \omega(|y(T)|, |y(T)|) \leqq \bar{C}(|x|), \end{aligned}$$

for some $\bar{C}(\cdot)$. Then, let $\alpha_0 \in \Gamma^a[s, T]$ such that

$$\alpha_0[b(\cdot)](t) \equiv a \quad \forall b(\cdot) \in \bigcup_{b \in B} \mathcal{B}^{b,s}.$$

We have

$$V^{a,b}(s,x) \leqq \sup_{b(\cdot)\in\mathcal{B}^{b,s}} J^{a,b}_{s,x}(a, b(\cdot))$$
$$\leqq \sup_{b(\cdot)\in\mathcal{B}^{b,s}} \left[I - \sum_{j\geqq 1} l(\tau_j, b_{j-1}, b_j) \right] \leqq \bar{C}(|x|).$$

On the other hand, for any $\alpha \in \Gamma^a[s, T]$, let $b_0(t) \equiv b$. Then, we have

$$V^{a,b}(s,x) \geqq \inf_{\alpha\in\Gamma^a[s,T]} J^{a,b}_{s,x}(\alpha[b_0(\cdot)], b_0(\cdot))$$
$$\geqq \inf_{\alpha\in\Gamma^a[s,T]} \left[-I + \sum_{i\geqq 1} k(\theta_j, a_{i-1}, a_i) \right] \geqq -\bar{C}(|x|).$$

Similarly, we can get the same thing for $U^{a,b}(s,x)$. This proves (2.18). Finally, let us prove the Lipschitz continuity of the lower and upper value functions in the time variable s. We still concentrate on the lower value function $V^{a,b}(s,x)$. Let $0 \leqq s \leqq \bar{s} \leqq T$. Then, for any

$$a(\cdot)\in\mathcal{A}^{a,s}, \qquad \hat{a}(\cdot)\in\mathcal{A}^{a,\bar{s}},$$
$$b(\cdot)\in\mathcal{B}^{b,s}, \qquad \hat{b}(\cdot)\in\mathcal{B}^{b,\bar{s}},$$

with

$$a(\cdot)|_{[\bar{s},T]} = \hat{a}(\cdot), \qquad b(\cdot)|_{[\bar{s},T]} = \hat{b}(\cdot),$$

and any $t\in[\bar{s}, T]$, we have

$$|y_{s,x}(t) - y_{\bar{s},x}(t)| \leqq \int_s^{\bar{s}} |g(\tau, y_{s,x}(\tau), a(\tau), b(\tau))|\, d\tau$$
$$+ \int_{\bar{s}}^t |g(\tau, y_{s,x}(\tau), a(\tau), b(\tau)) - g(\tau, y_{\bar{s},x}(\tau), \hat{a}(\tau), \hat{b}(\tau))|\, d\tau$$
$$\leqq \int_s^{\bar{s}} L(1+|y_{s,x}(\tau)|)\, d\tau + \int_{\bar{s}}^T L|y_{s,x}(\tau) - y_{\bar{s},x}(\tau)|\, d\tau.$$

Hence, by the Gronwall inequality, we have

$$|y_{s,x}(t) - y_{\bar{s},x}(t)| \leqq C(1+|x|)(\bar{s}-s) \quad \forall t\in[\bar{s}, T],$$

where C is some constant. Now, for any $b(\cdot)\in\mathcal{B}^{b,s}$ and $\hat{\alpha}\in\Gamma^a[\bar{s}, T]$, we define

$$\hat{b}(t) = b(t), \quad t\geqq\bar{s}, \qquad \hat{b}(\bar{s}-0) = b,$$

and

$$\alpha[b(\cdot)](t) = \begin{cases} a, & t\in[s,\bar{s}), \\ \hat{\alpha}[\hat{b}(\cdot)](t), & t\in[\bar{s}, T]. \end{cases}$$

Then, we have

$$J^{a,b}_{s,x}(\alpha[b(\cdot)], b(\cdot)) = \int_s^{\bar{s}} f(t, y_{s,x}(t), \alpha[b(\cdot)](t), b(t))\, dt$$
$$+ J^{a,b}_{\bar{s},x}(\hat{\alpha}[\hat{b}(\cdot)], \hat{b}(\cdot)) - \sum_{\tau_j\leqq\bar{s}} l(\tau_j, b_{j-1}, b_j)$$
$$+ \int_{\bar{s}}^T [f(t, y_{s,x}(t), \alpha[b(\cdot)](t), b(t))$$
$$- f(t, y_{\bar{s},x}(t), \alpha[b(\cdot)](t), b(t))]\, dt$$
$$\leqq C(|x|)(\bar{s}-s) + J^{a,b}_{\bar{s},x}(\hat{\alpha}[\hat{b}(\cdot)], \hat{b}(\cdot)).$$

Thus,

$$\sup_{b(\cdot)\in\mathcal{B}^{b,s}} J^{a,b}_{s,x}(\alpha[b(\cdot)], b(\cdot)) \leqq \sup_{\hat{b}(\cdot)\in\mathcal{B}^{b,\bar{s}}} J^{a,b}_{\bar{s},x}(\hat{\alpha}[\hat{b}(\cdot)], \hat{b}(\cdot)) + C(|x|)(\bar{s}-s).$$

Hence,

$$V^{a,b}(s,x) \leqq V^{a,b}(\bar{s},x) + C(|x|)(\bar{s}-s).$$

Conversely, for any $\hat{b}(\cdot)\in\mathcal{B}^{b,\bar{s}}$ and $\alpha\in\Gamma^a[s,T]$, we define

$$b(t) = \begin{cases} b, & t\in[s,\bar{s}), \\ \hat{b}(t), & t\in[\bar{s},T], \end{cases}$$

and

$$\hat{\alpha}[\hat{b}(\cdot)](t) = \alpha[b(\cdot)](t), \qquad t\in[\bar{s},T],$$
$$\hat{\alpha}[\hat{b}(\cdot)](\bar{s}-0) = a.$$

Then, we see that $b(\cdot)\in\mathcal{B}^{b,s}$ and $\hat{\alpha}\in\Gamma^a[\bar{s},T]$. It follows that

$$\begin{aligned} J^{a,b}_{s,x}(\alpha[b(\cdot)], b(\cdot)) &\geqq \int_s^{\bar{s}} f(t, y_{s,x}(t), \alpha[b(\cdot)](t), b(t))\,dt + J^{a,b}_{\bar{s},x}(\hat{\alpha}[\hat{b}(\cdot)], \hat{b}(\cdot)) \\ &\quad + \int_{\bar{s}}^{T} [f(t, y_{s,x}(t), \hat{\alpha}[\hat{b}(\cdot)](t), \hat{b}(t)) \\ &\quad - f(t, y_{\bar{s},x}(t), \hat{\alpha}[\hat{b}(\cdot)](t), \hat{b}(t))]\,dt \\ &\geqq -C(|x|)(\bar{s}-s) + J^{a,b}_{\bar{s},x}(\hat{\alpha}[\hat{b}(\cdot)], \hat{b}(\cdot)). \end{aligned}$$

Here, we have used (H2). Then, we see that

$$\begin{aligned} \sup_{b(\cdot)\in\mathcal{B}^{b,s}} J^{a,b}_{s,x}(\alpha[b(\cdot)], b(\cdot)) &\geqq \sup_{\hat{b}(\cdot)\in\mathcal{B}^{b,\bar{s}}} J^{a,b}_{\bar{s},x}(\alpha[\hat{b}(\cdot)], \hat{b}(\cdot)) - C(|x|)(\bar{s}-s) \\ &\geqq V^{a,b}(\bar{s},x) - C(|x|)(\bar{s}-s). \end{aligned}$$

Therefore,

$$V^{a,b}(s,x) \geqq V^{a,b}(\bar{s},x) - C(|x|)(\bar{s}-s).$$

We can obtain the same thing for the upper value function $U^{a,b}(s,x)$. Thus, (2.20) follows and we complete the proof. □

3. Optimality conditions, Isaacs' equation. In this section, we use the Bellman dynamic programming principle to derive the Isaacs equations for the lower and upper value functions of our game.

THEOREM 3.1. *The lower value function $V(\cdot,\cdot)$ satisfies the following optimality condition. For any $(a,b)\in A\times B$, $x\in X$, and $0\leqq s<\bar{s}\leqq T$,*

$$\begin{aligned} V^{a,b}(s,x) = \inf_{\alpha\in\Gamma^a[s,T]} \sup_{b(\cdot)\in\mathcal{B}^{b,s}} \Big\{ &\int_s^{\bar{s}} f(t, y_{s,x}(t), \alpha[b(\cdot)](t), b(t))\,dt \\ &+ \sum_{\theta_i\leqq\bar{s}} k(\theta_i, a_{i-1}, a_i) - \sum_{\tau_j\leqq\bar{s}} l(\tau_j, b_{j-1}, b_j) \\ &+ V^{\alpha[b(\cdot)](\bar{s}),b(\bar{s})}(\bar{s}, y_{s,x}(\bar{s})) \Big\}, \end{aligned} \tag{3.1}$$

where $\{a_i,\theta_i\}$ and $\{b_j,\tau_j\}$ are associated with $\alpha[b(\cdot)]$ and $b(\cdot)$, respectively, and $\alpha[b(\cdot)](\bar{s}) = \alpha[b(\cdot)](\bar{s}+0)$, $b(\bar{s}) = b(\bar{s}+0)$.

The proof of this theorem is very similar to that given in [9] and [22]. For the convenience of the readers, we give a sketch of the proof.

Sketch of the proof. Let $w(s, x)$ be the right-hand side of (3.1). For any $\varepsilon > 0$, there exists an $\alpha_0 \in \Gamma^a[s, T]$, such that

$$\begin{aligned} w(s,x)+\varepsilon \geqq \sup_{b(\cdot)\in\mathcal{B}^{b,s}} \Big\{ \int_s^{\bar{s}} & f(t, y_{s,x}(t), \alpha_0[b(\cdot)](t), b(t))\, dt \\ & + \sum_{\theta_i \leqq \bar{s}} k(\theta_i, a_{i-1}, a_i) - \sum_{\tau_j \leqq \bar{s}} l(\tau_j, b_{j-1}, b_j) \\ & + V^{\alpha_0[b(\cdot)](\bar{s}), b(\bar{s})}(\bar{s}, y_{s,x}(\bar{s})) \Big\}. \end{aligned} \tag{3.2}$$

On the other hand, we have $\hat{\alpha} \equiv \hat{\alpha}(y_{s,x}(\bar{s}), \alpha_0[b(\cdot)](\bar{s}), b(\bar{s})) \in \Gamma^{\alpha_0[b(\cdot)](\bar{s})}[\bar{s}, T]$, such that

$$V^{\alpha_0[b(\cdot)](\bar{s}), b(\bar{s})}(\bar{s}, y_{s,x}(\bar{s})) + \varepsilon \geqq \sup_{\hat{b}(\cdot)\in\mathcal{B}^{b(\bar{s}),\bar{s}}} J^{\alpha_0[b(\cdot)](\bar{s}), b(\bar{s})}_{\bar{s}, y_{s,x}(\bar{s})}(\hat{\alpha}[\hat{b}(\cdot)], \hat{b}(\cdot)). \tag{3.3}$$

Thus, if we define $\bar{\alpha} \in \Gamma^a[s, T]$ by the following: for any $b(\cdot) \in \bigcup_{b\in B} \mathcal{B}^{b,s}$,

$$\bar{\alpha}[b(\cdot)](t) = \begin{cases} \alpha_0[b(\cdot)](t), & t \in [s, \bar{s}], \\ \hat{\alpha}[b(\cdot)]_{[\bar{s},T]}(t), & t \in (\bar{s}, T], \end{cases} \tag{3.4}$$

then, under this strategy, we can obtain with some calculation that

$$\sup_{b(\cdot)\in\mathcal{B}^{b,s}} J^{a,b}_{s,x}(\bar{\alpha}[b(\cdot)], b(\cdot)) \leqq w(s,x) + 2\varepsilon.$$

Hence, we see that

$$V^{a,b}(s,x) \leqq w(s,x). \tag{3.5}$$

Conversely, for any $\varepsilon > 0$, we have an $\alpha_1 \in \Gamma^a[s, T]$, such that

$$V^{a,b}(s,x) + \varepsilon \geqq \sup_{b(\cdot)\in\mathcal{B}^{b,s}} J^{a,b}_{s,x}(\alpha_1[b(\cdot)], b(\cdot)). \tag{3.6}$$

On the other hand, by the definition of $w(s, x)$, there exists a $b_1(\cdot) \in \mathcal{B}^{b,s}$, such that

$$\begin{aligned} w(s,x) \leqq & \int_s^{\bar{s}} f(t, y_{s,x}(t), \alpha_1[b_1(\cdot)](t), b_1(t))\, dt + \sum_{\theta_i \leqq \bar{s}} k(\theta_i, a_{i-1}, a_i) \\ & - \sum_{\tau_j \leqq \bar{s}} l(\tau_j, b_{j-1}, b_j) + V^{\alpha_1[b_1(\cdot)](\bar{s}), b_1(\bar{s})}(\bar{s}, y_{s,x}(\bar{s})) + \varepsilon. \end{aligned} \tag{3.7}$$

Now, for any $\tilde{b}(\cdot) \in \mathcal{B}^{b_1(\bar{s}),\bar{s}}$, we define

$$\hat{b}(t) = b_1(t)\chi_{[s,\bar{s}]}(t) + \tilde{b}(t)\chi_{[\bar{s},T]}(t) \tag{3.8}$$

and define $\hat{\alpha} \in \Gamma^{\alpha_1[b_1(\cdot)](\bar{s})}[\bar{s}, T]$ as follows:

$$\hat{\alpha}[\tilde{b}(\cdot)](t) = \alpha_1[\hat{b}(\cdot)](t), \qquad t \in [\bar{s}, T]. \tag{3.9}$$

Then, under this $\hat{\alpha}$, we have

$$V^{\alpha_1[b_1(\cdot)](\bar{s}), b_1(\bar{s})}(\bar{s}, y_{s,x}(\bar{s})) \leqq \sup_{\tilde{b}(\cdot)\in\mathcal{B}^{b_1(\bar{s}),\bar{s}}} J^{\alpha_1[b_1(\cdot)](\bar{s}), b_1(\bar{s})}_{\bar{s}, y_{s,x}(\bar{s})}(\hat{\alpha}[\tilde{b}(\cdot)], \tilde{b}(\cdot)). \tag{3.10}$$

Hence, there exists a $\tilde{b}_1(\cdot) \in \mathcal{B}^{b_1(\bar{s}),\bar{s}}$, such that

$$V^{\alpha_1[b_1(\cdot)](\bar{s}), b_1(\bar{s})}(\bar{s}, y_{s,x}(\bar{s})) \leqq J^{\alpha_1[b_1(\cdot)](\bar{s}), b_1(\bar{s})}_{\bar{s}, y_{s,x}(\bar{s})}(\hat{\alpha}[\tilde{b}_1(\cdot)], \tilde{b}_1(\cdot)) + \varepsilon. \tag{3.11}$$

Thus, by defining

$$\hat{b}_1(t) = b_1(t)\chi_{[s,\bar{s}]}(t) + \tilde{b}_1(t)\chi_{[\bar{s},T]}(t)$$

we have

$$\begin{aligned} w(s,x)-\varepsilon &\leqq \int_s^{\bar{s}} f(t, y_{s,x}(t), \alpha_1[b_1(\cdot)](t), b_1(t))\,dt + \sum_{\theta_i \leqq \bar{s}} k(\theta_i, a_{i-1}, a_i) \\ &\quad - \sum_{\tau_j \leqq \bar{s}} l(\tau_j, b_{j-1}, b_j) + J_{\bar{s}, y_{s,x}(\bar{s})}^{\alpha_1[b_1(\cdot)](\bar{s}), b_1(\bar{s})}(\hat{\alpha}[\tilde{b}_1(\cdot)], \tilde{b}_1(\cdot)) + \varepsilon \\ &= J_{s,x}^{a,b}(\hat{\alpha}_1[\hat{b}_1(\cdot)], \hat{b}_1(\cdot)) + \varepsilon \\ &\leqq V^{a,b}(s,x) + 2\varepsilon. \end{aligned}$$

Let $\varepsilon \to 0$, and we get the conclusion. □

Next, let us define the following mappings. For any $(m \times n)$ matrix-valued function $W(\cdot,\cdot) = (W^{a,b}(\cdot,\cdot))$ defined on $[0, T] \times X$,

$$M^{a,b}[W](s,x) = \min_{\bar{a} \neq a} \{W^{\bar{a},b}(s,x) + k(s,a,\bar{a})\},$$

$$M_{a,b}[W](s,x) = \max_{\bar{b} \neq b} \{W^{a,\bar{b}}(s,x) - l(b,\bar{b})\}.$$

These two mappings are called obstacle operators. From Theorem 3.1 above, we can obtain the following theorem.

THEOREM 3.2. *The lower value function $V(\cdot)$ satisfies the following*:

(i) *For any $(a, b, s, x) \in A \times B \times [0, T] \times X$,*

$$M_{a,b}[V](s,x) \leqq V^{a,b}(s,x) \leqq M^{a,b}[V](s,x). \tag{3.12}$$

(ii) *Suppose at $(a, b, s, x) \in A \times B \times [0, T] \times X$,*

$$V^{a,b}(s,x) < M^{a,b}[V](s,x). \tag{3.13}$$

Then, there exists an $s_0 > s$, such that for all $\bar{s} \in [s, s_0]$,

$$V^{a,b}(s,x) \geqq \int_s^{\bar{s}} f(s, y_{s,x}(s), a, b)\,ds + V^{a,b}(\bar{s}, y_{s,x}(\bar{s})). \tag{3.14}$$

(iii) *Suppose at $(a, b, s, x) \in A \times B \times [0, T] \times X$,*

$$V^{a,b}(s,x) > M_{a,b}[V](s,x). \tag{3.15}$$

Then, there exists an $s_0 > s$, such that for all $\bar{s} \in [s, s_0]$,

$$V^{a,b}(s,x) \leqq \int_s^{\bar{s}} f(s, y_{s,x}(s), a, b)\,ds + V^{a,b}(\bar{s}, y_{s,x}(\bar{s})). \tag{3.16}$$

Proof. As in [22] (see also [5]), we can obtain (i).

Now, let us prove (ii). By Theorem 3.1, we see that for any $\bar{s} > s$, $\varepsilon > 0$, and $b(\cdot) \equiv b \in B$, there exists an $\alpha_\varepsilon^{\bar{s}} \in \Gamma^a[s, T]$, such that

$$\begin{aligned} V^{a,b}(s,x) + \varepsilon \geqq &\int_s^{\bar{s}} f(t, y_{s,x}(t), \alpha_\varepsilon^{\bar{s}}[b](t), b)\,dt \\ &+ \sum_{\theta_i \leqq \bar{s}} k(\theta_i, a_{i-1}, a_i) + V^{\alpha_\varepsilon^{\bar{s}}[b](\bar{s}), b}(\bar{s}, y_{s,x}(\bar{s})), \end{aligned} \tag{3.17}$$

where $\alpha_\varepsilon^{\bar{s}}[b](\cdot) = \sum_{i \geqq 1} a_{i-1} \chi_{[\theta_{i-1}, \theta_i)}(\cdot) \in \mathscr{A}^{a,s}$. Then, we are ready to show that for all small $\varepsilon > 0$ and $\bar{s} > s$ with $\bar{s} - s$ sufficiently small,

$$\theta_1 \equiv \theta_1^{\varepsilon, \bar{s}} > \bar{s}. \tag{3.18}$$

Then, (3.17) becomes

$$V^{a,b}(s,x)+\varepsilon \geqq \int_s^{\bar{s}} f(t, y_{s,x}(t), a, b)\,ds + V^{a,b}(\bar{s}, y_{s,x}(\bar{s})). \tag{3.19}$$

Thus, fix an $\bar{s}>s$ with $\bar{s}-s$ small and let $\varepsilon \to 0$, and we obtain (ii).

To prove (iii), we let

$$\alpha_0[b(\cdot)](t) \equiv a \quad \forall b(\cdot) \in \bigcup_{b\in B} \mathcal{B}^{b,s}, \quad t\in[s,T].$$

Then we know that for any $\varepsilon>0$ and $\bar{s}\in[s,T]$, there exists a $b_\varepsilon^{\bar{s}}(\cdot)\in\mathcal{B}^{b,s}$ such that

$$\begin{aligned} V^{a,b}(s,x)-\varepsilon \leqq & \int_s^{\bar{s}} f(t, y_{s,x}(t), a, b_\varepsilon^{\bar{s}}(t))\,dt \\ & - \sum_{\tau_j \leqq \bar{s}} l(\tau_j, b_{j-1}, b_j) + V^{a, b_\varepsilon^{\bar{s}}(\bar{s})}(\bar{s}, y_{s,x}(\bar{s})); \end{aligned} \tag{3.20}$$

here, $b_\varepsilon^{\bar{s}}(\cdot)=\sum_{j\geqq 1} b_{j-1}\chi_{[\tau_{j-1},\tau_j)}(\cdot)$. Then, as above, we are able to prove

$$\tau_1 \equiv \tau_1^{\varepsilon,\bar{s}} > \bar{s},$$

for all $\varepsilon>0$ and $\bar{s}>s$ with $\bar{s}-s$ small enough. Then, our conclusion (iii) follows. □

Now, we can derive the Isaacs' equation for the lower value function $V(\cdot,\cdot)$. Let

$$H^{a,b}(s,x,p)=\langle p, g(s,x,a,b)\rangle + f(s,x,a,b). \tag{3.21}$$

THEOREM 3.3. *Suppose $V(\cdot,\cdot)$ is a C^1 $(m\times n)$-matrix-valued function. Then, for any $(a,b)\in A\times B$,*

$$M_{a,b}[V](s,x) \leqq V^{a,b}(s,x) \leqq M^{a,b}[V](s,x) \quad \forall (s,x)\in[0,T]\times X; \tag{3.22}$$

on the set $\{(s,x)\in[0,T]\times X \mid M_{a,b}[V](s,x) < V^{a,b}(s,x)\}$,

$$V_s^{a,b}(s,x)+H^{a,b}(s,x,V_x^{a,b}(s,x)) \geqq 0; \tag{3.23}$$

and on the set $\{(s,x)\in[0,T]\times X \mid M^{a,b}[V](s,x) > V^{a,b}(s,x)\}$,

$$V_s^{a,b}(s,x)+H^{a,b}(s,x,V_x^{a,b}(s,x)) \leqq 0. \tag{3.24}$$

The terminal condition is given by

$$V^{a,b}(T,x)=h(x). \tag{3.25}$$

Proof. We can apply Theorem 3.2 directly. □

We see that (3.22)–(3.25) is a system of evolutionary quasi-variational inequalities ([2]) with bilateral obstacles. This is the corresponding Isaacs equation that the lower value function $V(\cdot,\cdot)$ should satisfy (in some sense).

THEOREM 3.4. *Suppose $V(\cdot,\cdot)$ is a C^1 $(m\times n)$-matrix-valued function. Then, it solves* (3.22)–(3.25) *if and only if it satisfies the following two systems*:

$$\begin{aligned} &\max\{\min\{V_s^{a,b}(x)+H^{a,b}(s,x,V_x^{a,b}(s,x)), M^{a,b}[V](s,x)-V^{a,b}(s,x)\}, \\ &\qquad M_{a,b}[V](s,x)-V^{a,b}(s,x)\}=0, \\ &\qquad (a,b,s,x)\in A\times B\times[0,T)\times X, \\ &V^{a,b}(T,x)=h(x) \quad \forall (a,b,x)\in A\times B\times X \end{aligned} \tag{3.26}$$

and

$$\begin{aligned}&\min\{\max\{V_s^{a,b}(s,x)+H^{a,b}(s,x,V_x^{a,b}(s,x)),\ M_{a,b}[V](x)-V^{a,b}(s,x)\},\\ &\qquad M^{a,b}[V](x)-V^{a,b}(s,x)\}=0,\\ &\qquad (a,b,s,x)\in A\times B\times[0,T]\times X,\\ &V^{a,b}(T,x)=h(x)\quad \forall (a,b,x)\in A\times B\times X.\end{aligned}\tag{3.27}$$

The proof follows from some straightforward computations involving min and max.

Remark 3.5. In [22] such a simple equivalence was not pointed out. We will return to this point in Remark 4.5.

Symmetrically, for the upper value function $U(\cdot,\cdot)$, we have the following dynamic programming principle.

THEOREM 3.6. *The function $U(\cdot,\cdot)$ satisfies the following optimality principle. For all $(a,b,s,x)\in A\times B\times[0,T]\times X$, $\bar{s}>s$,*

$$\begin{aligned}U^{a,b}(s,x)=\sup_{\beta\in\Delta^b[s,T]}\inf_{a(\cdot)\in\mathcal{A}^{a,s}}\Big\{&\int_s^{\bar{s}} f(t,y_{s,x}(t),a(t),\beta[a(\cdot)](t))\,dt\\ &+\sum_{\theta_i\leqq\bar{s}} k(\theta_i,a_{i-1},a_i)-\sum_{\tau_j\leqq\bar{s}} l(\tau_j,b_{j-1},b_j)\\ &+U^{a(\bar{s}),\beta[a(\cdot)](\bar{s})}(\bar{s},y_{s,x}(\bar{s}))\Big\},\end{aligned}\tag{3.28}$$

where $\{a_i,\theta_i\}$ and $\{b_j,\tau_j\}$ are associated with $a(\cdot)$ and $\beta[a(\cdot)]$, respectively.

It is important to know that from the theorem above, we can obtain exactly the *same* result as in Theorem 3.2 (therefore exactly the *same* results as in Theorems 3.3 and 3.4) for the upper value function $U(\cdot,\cdot)$. Hence, we have the following proposition.

PROPOSITION 3.7. *If* (3.22)–(3.25) *admit at most one C^1 solution and $U(\cdot,\cdot)$ and $V(\cdot,\cdot)$ are C^1, then,*

$$U(\cdot,\cdot)=V(\cdot,\cdot).\tag{3.29}$$

That is, the game has a (Elliot-Kalton) value.

Unfortunately, it is well known that the upper and the lower value functions are not necessarily C^1 and, in a similar manner to a usual first-order Hamilton-Jacobi-Bellman equation, the problem (3.22)–(3.25) may have no C^1 solutions. Thus, Proposition 3.7 actually does not tell us much. We need additional investigations.

4. Uniqueness of viscosity solutions, existence of the value. In this section, we introduce some generalized notion of solutions to the Isaacs equations (3.22)–(3.25). This notion was introduced by Crandall and Lions [7] (see also [6], [9], [15]–[17], [19], [21]–[23]). Let us start with (3.26) and (3.27).

DEFINITION 4.1. Function $W(\cdot,\cdot)\in C([0,T]\times X;\mathbb{R}^{m\times n})$ is called a viscosity supersolution (subsolution) of problem (3.26) if $W^{a,b}(T,x)=h(x)$ for all $(a,b,x)\in A\times B\times X$ and if $\varphi\in C^1$ with $W^{a,b}-\varphi$ attains a local maximum (minimum) at $(t_0,x_0)\in[0,T)\times X$, then

$$\begin{aligned}&\max\{\min\{\varphi_s(x)+H^{a,b}(s,x,\varphi_x(s,x)),\ M^{a,b}[W](s,x)-W^{a,b}(s,x)\},\\ &\qquad M_{a,b}[W](s,x)-W^{a,b}(s,x)\}\geqq 0(\leqq 0).\end{aligned}\tag{4.1}$$

Here, if $t_0=0$, then, $\varphi_s(0,x_0)$ is understood as the right derivative. If $W(\cdot)$ is both a viscosity sub- and supersolution of (3.26), then it is called a viscosity solution of (3.26).

In the same manner, we can define the viscosity subsolutions, supersolutions, and solutions for (3.27). It is not hard to see that [5] in the above definition, we can replace the local maximum (respectively, minimum) by a strict local maximum (respectively,

minimum). Also, we note that the map $H^{a,b}(s, x, p)$ does not have to be of the form (3.21). Actually, in this section, we make the following assumption concerning this map (we call this assumption (H1′) because it is a replacement for (H1)):

(H1′) There exist a constant $L>0$ and a nondecreasing function $\omega_r(\cdot):[0,\infty)\to[0,\infty)$ with $\omega_r(0)=0$ for all $r\geqq 0$, such that, for all $(a, b, s)\in A\times B\times[0, T]$,

$$|H^{a,b}(s, x, p)-H^{a,b}(s, x, q)|\leqq L(1+|x|)|p-q| \quad \forall x, p, q\in X, \tag{4.2}$$

$$|H^{a,b}(s, x, p)-H^{a,b}(s, y, q)|\leqq \omega_r(|x-y|(1+|p|)) \quad \forall x, y, p\in X, \quad |x|, |y|\leqq r. \tag{4.3}$$

It is easy to see that (H1) implies (H1′).

Next, from Theorem 3.3, it is reasonable to give the following definition.

DEFINITION 4.2. A continuous function $W(\cdot,\cdot):[0, T]\times X\to\mathbb{R}^{m\times n}$ is called a viscosity solution of problem (3.22)-(3.25), if it is a viscosity solution of both (3.26) and (3.27).

THEOREM 4.3. *The lower value function $V(\cdot)$ and the upper value function $U(\cdot)$ are viscosity solutions of* (3.22)-(3.25).

The proof immediately follows Theorem 3.2 and Definitions 4.1 and 4.2. The rest of this section is devoted to showing the existence of the value for our differential game. To this end, we first give the following lemma.

LEMMA 4.4. *Let $W(\cdot,\cdot)$ and $\hat{W}(\cdot,\cdot)$ be viscosity subsolution and supersolution of* (3.26) *and* (3.27), *respectively. Then, for all $(a, b, s, x)\in A\times B\times[0, T]\times X$,*

$$W^{a,b}(s, x)\geqq M_{a,b}[W](s, x), \tag{4.4}$$

$$\hat{W}^{a,b}(s, x)\leqq M^{a,b}[\hat{W}](s, x). \tag{4.5}$$

The proof is almost clear by applying the argument given in [5].

Remark 4.5. We should note that if $W(\cdot,\cdot)$ is only a viscosity solution of (3.26) (or (3.27)), then it is not clear whether it satisfies

$$M_{a,b}[W](s, x)\leqq W^{a,b}(s, x)\leqq M^{a,b}[W](s, x) \quad \forall (a, b, s, x)\in A\times B\times[0, T]\times X. \tag{4.6}$$

There was a careless mistake concerning this matter in [22]. Fortunately, by applying a similar argument used in this paper, the final result of [22], i.e., the existence of the value for the differential game, remains true.

Now, let us make a further assumption that will play a very important role in proving the next theorem.

(H4) For any loop $\{(a_i, b_i)\}_{i=1}^{j}\subset A\times B$, with the properties that

$$\begin{aligned} &j\leqq mn, \quad a_{j+1}=a_1, \quad b_{j+1}=b_1;\\ &\text{either} \quad a_{i+1}=a_i, \quad \text{or } b_{i+1}=b_i \quad \forall 1\leqq i\leqq j, \end{aligned} \tag{4.7}$$

it holds that

$$\sum_{i=1}^{j} k(s, a_i, a_{i+1})-\sum_{i=1}^{j} l(s, b_i, b_{i+1})\neq 0 \quad \forall s\in[0, T]. \tag{4.8}$$

Now, we can state and prove the following important result.

THEOREM 4.6. *Suppose* (H1′) *and* (H2)-(H4) *hold. Let $W(\cdot,\cdot)$ and $\hat{W}(\cdot,\cdot)$ be two viscosity solutions of* (3.22)-(3.25). *Then,*

$$W^{a,b}(s, x)=\hat{W}^{a,b}(s, x) \quad \forall (a, b, s, x)\in A\times B\times[0, T]\times X. \tag{4.9}$$

Proof. We prove the theorem by contradiction. Thus, we may assume that

$$\max_{(a,b)\in A\times B}\ \sup_{(t,x)\in\mathcal{O}}[W^{a,b}(t, x)-\hat{W}^{a,b}(t, x)]\geqq\bar{\sigma}>0, \tag{4.10}$$

where

$$\mathcal{O}=\{(t,x)\in(T-T_0,T)\times X\,|\,|x|<L_0(t-T+T_0)\},$$

with

$$0\leqq T_0<T,\quad T_0<\frac{1}{L},\quad L_0=\frac{l}{1-LT_0}.$$

Then, by (4.2), for all $(t,x)\in\mathcal{O}$ and $p,q\in X$,

$$|H^{a,b}(t,x,p)-H^{a,b}(t,x,q)|\leqq L_0|p-q|. \tag{4.11}$$

Let $\varepsilon,\delta>0$, with $\varepsilon+\delta<L_0T_0$ and let $K>0$, $\zeta\in C^\infty(\mathbb{R})$, with the properties that

$$K>\sup\{|W^{a,b}(t,x)-\hat{W}^{a,b}(s,y)|\,|\,(t,x,s,y)\in\mathcal{O}^2,(a,b)\in A\times B\}, \tag{4.12}$$

$$\zeta(r)=\begin{cases}0, & r\leqq-\delta,\\ -K, & r\geqq 0,\end{cases} \tag{4.13}$$

$$\zeta'(r)\leqq 0\quad\forall r\in\mathbb{R}. \tag{4.14}$$

Then, for $\alpha,\beta>0$ and $\sigma\geqq 0$, we define

$$\Psi_0^{a,b}(t,x)=W^{a,b}(t,x)-\hat{W}^{a,b}(t,x)+2\zeta(\langle x\rangle_\varepsilon-L_0(t-T+T_0))+2\sigma(t-T) \tag{4.15}$$

where $\langle x\rangle_\varepsilon=(|x|^2+\varepsilon^2)^{1/2}$. As the first step of the proof, we have the following lemma.

LEMMA 4.7. *There exist* $\varepsilon_0,\delta_0,\sigma_0,\gamma_0>0$, *such that for any* $0<\varepsilon\leqq\varepsilon_0$, $0<\delta\leqq\delta_0$, $0<\sigma\leqq\sigma_0$, *there exist* $(a_0,b_0)\in A\times B$ *and* $(t_0,x_0)\in\mathcal{O}$, *such that*

$$\Psi_0^{a_0,b_0}(t_0,x_0)=\max_{(a,b)\in A\times B}\max_{(t,x)\in\bar{\mathcal{O}}}\Psi_0^{a,b}(t,x), \tag{4.16}$$

$$\langle x_0\rangle_\varepsilon<L_0(t_0-T+T_0), \tag{4.17}$$

$$t_0\leqq T-\gamma_0, \tag{4.18}$$

$$\begin{aligned}&W^{a_0,b_0}(t_0,x_0)>M_{a_0,b_0}[W](t_0,x_0),\\ &\hat{W}^{a_0,b_0}(t_0,x_0)<M^{a_0,b_0}[\hat{W}](t_0,x_0).\end{aligned} \tag{4.19}$$

The proof follows from the ideas of [18] (see also [14], [22], [23]) and (H4).

Next, for any $\alpha,\beta>0$, we define

$$\begin{aligned}\Psi(t,x,s,y)=&W^{a_0,b_0}(t,x)-\hat{W}^{a_0,b_0}(s,y)-\frac{1}{\alpha}|x-y|^2-\frac{1}{\beta}|t-s|^2\\ &+\zeta(\langle x\rangle_\varepsilon-L_0(t-T+T_0))+\zeta(\langle y\rangle_\varepsilon-L_0(s-T+T_0))+\sigma(t+s)-2\sigma T\\ &\forall(t,x,s,y)\in\bar{\mathcal{O}}^2.\end{aligned} \tag{4.20}$$

Here, (a_0,b_0) is obtained from Lemma 4.7. Thus, it depends on ε, δ, σ in general. For this function, we have the following Lemma.

LEMMA 4.8. *For any* $\hat{\varepsilon}>0$, *there exist* $\hat{\alpha},\hat{\beta}>0$, *such that for all* $0<\alpha\leqq\hat{\alpha}$ *and* $0<\beta\leqq\hat{\beta}$,

$$\max_{\bar{\mathcal{O}}^2}\Psi(t,x,s,y)=\max_{\bar{\mathcal{O}}}\Psi_0^{a,b}(t,x)+\hat{\varepsilon}. \tag{4.21}$$

The proof is straightforward by using the ideas of [7] and [12].

Now, we are able to complete the proof of Theorem 4.6 by combining Lemmas 4.4, 4.7, and 4.8. In fact, by Lemma 4.4, we know that there exists a $\hat{\delta}>0$, such that

$$\hat{\mathcal{O}}=\{(t,x)\in(T-T_0,T)\times X\,|\,|t-t_0|^2+|x-x_0|^2<\hat{\delta}^2\}\subset\mathcal{O}, \tag{4.22}$$

and for all $(t, x) \in \hat{\mathcal{O}}$,

$$
\begin{aligned}
&W^{a_0,b_0}(t, x) > M_{a_0,b_0}[W](t, x), \\
&\hat{W}^{a_0,b_0}(t, x) < M^{a_0,b_0}[\hat{W}](t, x).
\end{aligned} \tag{4.23}
$$

We note that $\hat{\delta}$ only depends on ε, δ, and γ_0. Then, an argument similar to that given in [7] and [12] will give

$$\sigma \leqq 0,$$

which is a contradiction because we have assumed that $\sigma > 0$. This completes the proof of the theorem. □

Remark 4.9. In the proof of the theorem above, it is not hard to find that the technique we used in [16] does not apply here due to the nature of the game. However, the method we used here is applicable to the problem studied in [16]. Actually, by the argument used here, we can simplify the proof of the relevant result in [16]. Also, the proof will be much easier if the switching cost functions k and l are independent of the time variable t.

Now, by Theorems 4.3 and 4.6, we can obtain the following theorem.

THEOREM 4.10. *Let* (H1)-(H4) *hold. then the Elliot-Kalton value of our differential game* (1.1)-(1.2) *exists.*

5. A limiting case. It is interesting to investigate what happens if the switching costs $k(\cdot, \cdot, \cdot)$ and $l(\cdot, \cdot, \cdot)$ approach zero. Let us study such a situation in this section. We let k_ε and l_ε be two sequences of switching costs for players I and II, respectively, with the properties that

$$\lim_{\varepsilon \to 0} k_\varepsilon(t, a, \hat{a}) = 0 \quad \forall t \in [0, T], \quad a, \hat{a} \in A, \tag{5.1}$$

$$\lim_{\varepsilon \to 0} l_\varepsilon(t, b, \hat{b}) = 0 \quad \forall t \in [0, T], \quad b, \hat{b} \in B. \tag{5.2}$$

We let (H1)-(H3) hold. From Lemma 2.3, we see that the family of the lower value functions (denoted by) $V_\varepsilon(\cdot, \cdot)$ corresponding to switching costs $(k_\varepsilon, l_\varepsilon)(\varepsilon > 0)$ is uniformly bounded and equicontinuous in bounded sets. Thus, by the Arzela-Ascoli theorem, we can find a subsequence (still denoted by) $V_\varepsilon(\cdot, \cdot)$, such that

$$\lim_{\varepsilon \to 0} V_\varepsilon^{a,b}(t, x) = W^{a,b}(t, x), \tag{5.3}$$

uniformly for $t \in [0, T]$ and x in bounded sets. It is clear that $W^{a,b}(\cdot, \cdot)$ also satisfies (2.18)-(2.20).

Next, we will discuss further properties of $W^{a,b}(\cdot, \cdot)$. To this end, let us first define the following maps:

$$H^+(t, x, p) = \min_{a \in A} \max_{b \in B} \{\langle p, g(t, x, a, b)\rangle + f(t, x, a, b)\} \quad \forall (t, x, p) \in [0, T] \times X \times X, \tag{5.4}$$

$$H^-(t, x, p) = \max_{b \in B} \min_{a \in A} \{\langle p, g(t, x, a, b)\rangle + f(t, x, a, b)\} \quad \forall (t, x, p) \in [0, T] \times X \times X. \tag{5.5}$$

Then, our main result of this section can be stated as follows.

THEOREM 5.1. *Let* (H1)-(H3) *hold. Let* $\{W^{a,b}(\cdot, \cdot) | (a, b) \in A \times B\}$ *be any functions obtained through* (5.3). *Then, we have the following conclusions*:

(i) *There exists a function* $w(\cdot)$ *satisfying* (2.18)-(2.20), *such that*

$$W^{a,b}(t, x) = w(t, x) \quad \forall (a, b, t, x) \in A \times B \times [0, T] \times X. \tag{5.6}$$

(ii) *Function $w(\cdot,\cdot)$ is a viscosity subsolution of the upper Isaacs equation*

$$\begin{aligned} &w_t(t,x)+H^+(t,x,w_x(t,x))=0, \qquad (t,x)\in[0,T)\times X,\\ &w(T,x)=h(x), \qquad x\in X. \end{aligned} \tag{5.7}$$

(iii) *Function $w(\cdot,\cdot)$ is a viscosity supersolution of the lower Isaacs equation*

$$\begin{aligned} &w_t(t,x)+H^-(t,x,w_x(t,x))=0, \qquad (t,x)\in[0,T)\times X,\\ &w(T,x)=h(x), \qquad x\in X. \end{aligned} \tag{5.8}$$

Proof. (i) From

$$M_{a,b}[V_\varepsilon](t,x)\leqq V_\varepsilon^{a,b}(t,x)\leqq M^{a,b}[V_\varepsilon](t,x),$$

by letting $\varepsilon\to 0$ along the subsequence in (5.3), we obtain (5.6).

(ii) Let $\varphi\in C^1([0,T)\times X)$ with $w-\varphi$ attain a strict local maximum at $(t_0,x_0)\in[0,T)\times X$. Since the convergence in (5.3) is uniformly in $t\in[0,t]$ and x in bounded sets, we see that for any $a\in A$, there exist $t_\varepsilon\to t_0$ and $x_\varepsilon\to x_0$, such that

$$\max_{b\in B} V_\varepsilon^{a,b}(t_\varepsilon,x_\varepsilon)-\varphi(t_\varepsilon,x_\varepsilon)>\max_{b\in B} V_\varepsilon^{a,b}(t,x)-\varphi(t,x) \quad \text{for } (t,x) \text{ near } (t_\varepsilon,x_\varepsilon). \tag{5.9}$$

We let $b_\varepsilon^a\in B$, such that

$$V_\varepsilon^{a,b_\varepsilon^a}(t_\varepsilon,x_\varepsilon)=\max_{b\in B} V_\varepsilon^{a,b}(t_\varepsilon,x_\varepsilon). \tag{5.10}$$

Then, noting for any $b\in B\backslash\{b_\varepsilon^a\}$, $l(b_\varepsilon^a,b)>0$, we must have

$$V_\varepsilon^{a,b_\varepsilon^a}(t_\varepsilon,x_\varepsilon)>M_{a,b_\varepsilon^a}[V_\varepsilon](t_\varepsilon,x_\varepsilon). \tag{5.11}$$

Thus, by the definition of viscosity solutions and (5.11), we obtain

$$\varphi_t(t_\varepsilon,x_\varepsilon)+H^{a,b_\varepsilon^a}(t_\varepsilon,x_\varepsilon,\varphi_x(t_\varepsilon,x_\varepsilon))\geqq 0. \tag{5.12}$$

Then, by choosing a subsequence if necessary, and taking the limits, we obtain

$$\varphi_t(t_0,x_0)+H^{a,\bar b}(t_0,x_0,\varphi_x(t_0,x_0))\geqq 0, \tag{5.13}$$

for some $\bar b\in B$ (depending on a, in general). Thus,

$$\min_{a\in A}\max_{b\in B}\{\varphi_t(t_0,x_0)+H^{a,b}(t_0,x_0,\varphi_x(t_0,x_0))\}\geqq 0, \tag{5.14}$$

i.e.,

$$\varphi_t(t_0,x_0)+H^+(t_0,x_0,\varphi_x(t_0,x_0))\geqq 0. \tag{5.15}$$

Finally, from (5.3), (2.14), and Lemma 2.3, we see that

$$w(T,x)=h(x) \quad \forall x\in X.$$

This proves (ii). The proof of (iii) is similar. □

From the above theorem we can obtain the following interesting result.

COROLLARY 5.2. *Let* (H1) *hold. Let the Isaacs condition hold*:

$$H^+(t,x,p)=H^-(t,x,p)\equiv H(t,x,p) \quad \forall (t,x,p)\in[0,T]\times X\times X. \tag{5.16}$$

Then, there exists a function $w(\cdot,\cdot)$ satisfying (2.18)-(2.20), *such that for any $(a,b)\in A\times B$,*

$$\lim_{\varepsilon\to 0} V_\varepsilon^{a,b}(t,x)=w(t,x), \tag{5.17}$$

uniformly for $t \in [0, T]$ and x in any bounded sets. Moreover, the function $w(\cdot, \cdot)$ is the unique viscosity solution of the following Isaacs equation:

$$\begin{aligned} &w_t(t, x) + H(t, x, w_x(t, x)) = 0, \qquad (t, x) \in [0, T) \times X, \\ &w(T, x) = h(x), \qquad x \in X. \end{aligned} \tag{5.18}$$

Proof. We need only note that the uniqueness of the viscosity solutions of (5.18) [6], [7], [9] implies the whole sequence $V_\varepsilon^{a,b}$ converges. □

It is not hard to see that $w(\cdot, \cdot)$ obtained in (5.17) is exactly the Elliot-Kalton value function of the classical two-player zero-sum differential game of fixed duration with control sets A and B. We also see that we have the same result as Corollary 5.2 for the upper value functions $U_\varepsilon^{a,b}(\cdot)$. Finally, we see that as far as the above convergence is concerned, condition (H4) is irrelevant.

6. Approximation of classical differential games. In this section, we consider the classical differential game

$$\begin{aligned} &\dot{y}(t) = g(t, y(t), a(t), b(t)), \qquad t \in (0, T], \\ &y(0) = x, \end{aligned} \tag{6.1}$$

with the payoff functional

$$J(a(\cdot), b(\cdot)) = \int_0^T f(t, y(t), a(t), b(t))\, dt + h(y(T)). \tag{6.2}$$

Here

$$a(\cdot) \in \mathscr{A} \equiv \{a(\cdot) : [0, T] \to A \mid a(\cdot) \text{ measurable}\},$$

and

$$b(\cdot) \in \mathscr{B} \equiv \{b(\cdot) : [0, T] \to B \mid b(\cdot) \text{ measurable}\}.$$

The sets A and B are closed subsets of some separable matric spaces (and thus they are not necessarily finite sets). We choose two sequences of sets $\{A_n\}$ and $\{B_n\}$, such that each of A_n and B_n contains exactly n points and the following hold:

$$A_n \subset A, \quad B_n \subset B \quad \forall n \geqq 1,$$

$$A_1 \subset A_2 \subset A_3 \subset \cdots, \qquad B_1 \subset B_2 \subset B_3 \subset \cdots,$$

and

$$\overline{\bigcup_{n=1}^{\infty} A_n} = A, \qquad \overline{\bigcup_{n=1}^{\infty} B_n} = B.$$

Also, we choose positive real numbers k_n and l_n, such that

$$k_n, l_n \to 0 \quad \text{as } n \to \infty$$

and for any n, k_n / l_n is an irrational number. Then, we consider the following approximating problem. We take (6.1) as the state equation and use A_n, B_n as control domains. The functional of the approximating game is taken to be

$$J(a(\cdot), b(\cdot)) = \int_0^T f(t, y(t), a(t), b(t))\, dt + h(y(T)) + N_{a(\cdot)} k_n - N_{b(\cdot)} l_n, \tag{6.3}$$

where $N_{a(\cdot)}$ is the total number of switchings made by $a(\cdot)$ on $[0, T]$, i.e.,

$$a(\cdot) = \sum_{i=1}^{N_{a(\cdot)}} a_{i-1}\chi_{[\theta_{i-1},\theta_i)}(\cdot),$$

and $N_{b(\cdot)}$ is the total number of switchings made by $b(\cdot)$. Then, by the results of § 4, the value of this approximating differential game exists. We let $V_n^{a,b}(\cdot,\cdot)$ be the value function corresponding to this game. It is clear that $\{V_n^{a,b}(\cdot,\cdot) \mid (a, b) \in A_n \times B_n,\ n \geqq 1\}$ is bounded and equicontinuous in bounded sets. Thus, for any fixed $(a, b) \in A_m \times B_m$, by choosing the appropriate subsequence (still denoted by) $V_n^{a,b}(\cdot,\cdot)$, we have

$$\lim_{n\to\infty} V_n^{a,b}(t, x) = v^{a,b}(t, x), \tag{6.4}$$

uniformly in $t \in [0, T]$ and x in any compact sets of X for some function $v^{a,b}(\cdot,\cdot)$. Since

$$V_n^{a,b}(t, x) \leqq \min_{\bar{a}\in A, \bar{a}\neq a} (V_n^{\bar{a},b}(t, x) + k_n),$$

it follows that for any given $\bar{a} \in A_m \backslash \{a\}$, we have

$$v^{a,b}(t, x) \leqq v^{\bar{a},b}(t, x) \quad \forall (t, x) \in [0, T] \times X.$$

Hence,

$$v^{a,b}(t, x) = v^{\bar{a},b}(t, x) \quad \forall (t, x) \in [0, T] \times X, \quad \bar{a}, a \in A_m.$$

Similarly, we have

$$v^{a,b}(t, x) = v^{a,\bar{b}}(t, x) \quad \forall (t, x) \in [0, T] \times X, \quad \bar{b}, b \in B_m.$$

Thus, we may let

$$v(t, x) = v^{a,b}(t, x) \quad \forall (t, x) \in [0, T] \times X, \quad (a, b) \in A_m \times B_m, \quad m \geqq 1. \tag{6.5}$$

Next, almost exactly the same as in § 5, we can prove that $v(\cdot,\cdot)$ is a viscosity subsolution of the upper Isaacs equation

$$\begin{aligned} &w_t(t, x) + H^+(t, x, w_x(t, x)) = 0, \qquad (t, x) \in [0, T) \times X, \\ &w(T, x) = h(x), \qquad x \in X, \end{aligned} \tag{6.6}$$

and a viscosity supersolution of the lower Isaacs equation

$$\begin{aligned} &w_t(t, x) + H^-(t, x, w_x(t, x)) = 0, \qquad (t, x) \in [0, T) \times X, \\ &w(T, x) = h(x), \qquad x \in X, \end{aligned} \tag{6.7}$$

where

$$H^+(t, x, p) = \inf_{a\in A} \sup_{b\in B} \{\langle p, g(t, x, a, b)\rangle + f(t, x, a, b)\} \quad \forall (t, x, p) \in [0, T] \times X \times X, \tag{6.8}$$

$$H^-(t, x, p) = \sup_{b\in B} \inf_{a\in A} \{\langle p, g(t, x, a, b)\rangle + f(t, x, a, b)\} \quad \forall (t, x, p) \in [0, T] \times X \times X. \tag{6.9}$$

Hence, we finally obtain the following theorem.

THEOREM 6.1. *Let* (H1) *and the Isaacs condition hold*:

$$H^+(t, x, p) = H^-(t, x, p) \quad \forall (t, x, p) \in [0, T] \times X \times X. \tag{6.10}$$

Then, the value function $v(\cdot,\cdot)$ *exists and can be approximated by the value functions* $V_n^{a,b}(\cdot,\cdot)$ *of the approximating games in the following sense*:

$$\lim_{n\to\infty} V_n^{a,b}(t,x) = v(t,x), \tag{6.11}$$

uniformly in $t\in[0,T]$ *and* x *in any compact sets of* X *and any* $(a,b)\in A_m\times B_m (m\geqq 1)$.

REFERENCES

[1] E. N. BARRON, L. C. EVANS, AND R. JENSEN, *Viscosity solutions of Isaacs' equations and differential games with Lipschitz controls*, J. Differential Equations, 53 (1984), pp. 213-233.

[2] A. BENSOUSSAN AND J. L. LIONS, *Impulse Control and Quasi-Variational Inequalities*, Bordes, Paris, 1984.

[3] L. D. BERKOVITZ, *The existence of value and saddle point in games of fixed duration*, SIAM J. Control Optim., 23 (1985), pp. 172-196.

[4] ———, *Characterizations of the values of differential games*, Appl. Math. Optim., 17 (1987), pp. 177-183.

[5] I. CAPUZZO DOLCETTA AND L. C. EVANS, *Optimal switching for ordinary differential equations*, SIAM J. Control Optim., 22 (1984), pp. 143-161.

[6] M. G. CRANDALL, L. C. EVANS, AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487-502.

[7] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[8] R. J. ELLIOT AND N. J. KALTON, *The existence of value in differential games*, in Mem. Amer. Math. Soc. 126, American Mathematical Society, Providence, RI, 1972.

[9] L. C. EVANS AND P. E. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773-797.

[10] W. H. FLEMING, *The convergence problem for differential games* II, Ann. of Math. Stud., 52 (1964), pp. 195-210.

[11] A. FRIEDMAN, *Differential Games*, John Wiley, New York, 1971.

[12] H. ISHII, *Uniqueness of unbounded viscosity solution of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 721-748.

[13] N. N. KRASOVSKII AND A. I. SUBBOTIN, *Game-Theoretical Control Problems*, Springer-Verlag, New York, 1988.

[14] S. M. LENHART AND N. YAMADA, *Switching control game for piecewise-deterministic processes and associated system of* PDEs, in Proc. 26th Annual IEEE Conference on Decision and Control, IEEE Computer Society, Washington, DC, 1987, pp. 1109-1110.

[15] P. L. LIONS AND P. E. SOUGANIDIS, *Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations*, SIAM J. Control Optim., 21 (1985), pp. 566-583.

[16] S. STOJANOVIC AND J. YONG, *Optimal switching for systems governed by nonlinear evolution equations*, Numer. Funct. Anal. Optim., 9, 10 (1987), pp. 995-1030.

[17] ———, *Optimal switching for partial differential equations* I, II, J. Math. Anal. Appl., 138 (1989), pp. 418-460.

[18] N. YAMADA, *A system of elliptic variational inequalities associated with a stochastic switching game*, Hiroshima Math. J., 13 (1983), pp. 109-132.

[19] ———, *Viscosity solutions for a system of elliptic variational inequalities with bilateral obstacles*, Funckcialaj Ekvacioj, 30 (1987), pp. 417-425.

[20] J. YONG, *On the Isaacs equation of differential games of fixed duration*, J. Optim. Theory Appl., 50 (1986), pp. 359-364.

[21] ———, *Systems governed by ordinary differential equations with continuous, switching and impulse controls*, Appl. Math. Optim., 20 (1989), pp. 223-236.

[22] ———, *Differential games with switching strategies*, J. Math. Anal. Appl., 145 (1990), pp. 455-469.

[23] ———, *Existence of the value for a differential game with switching strategies in a Banach space*, submitted.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1251-1269, November 1990

FINITE-DIMENSIONAL COMPENSATORS FOR INFINITE-DIMENSIONAL SYSTEMS VIA GALERKIN-TYPE APPROXIMATION*

KAZUFUMI ITO†

Abstract. In this paper existence and construction of stabilizing compensators for linear time-invariant systems defined on Hilbert spaces are discussed. An existence result is established using Galerkin-type approximations in which independent basis elements are used instead of the complete set of eigenvectors. A design procedure based on approximate solutions of the optimal regulator and optimal observer via Galerkin-type approximation is given and the Schumacher approach is used to reduce the dimension of compensators. A detailed discussion for parabolic and hereditary differential systems is included.

Key words. finite-dimensional compensators, infinite-dimensional systems, Riccati equations

AMS(MOS) subject classifications. 93B50, 93C25, 65J10

1. Introduction.

Consider the evolution equation on the Hilbert space Z

$$\frac{d}{dt}z(t) = Az(t) + Bu(t), \qquad z(0) = z_0 \in Z, \tag{1.1}$$

where $u(t)$ is a $\mathbb{R}^m$-valued control function, A is the infinitesimal generator of a strongly continuous semigroup $S(t)$ on Z, and $B \in L(\mathbb{R}^m, Z)$. The $\mathbb{R}^P$-valued observation function y is given by

$$y(t) = Cz(t), \qquad t \geqq 0. \tag{1.2}$$

We assume that $C \in L(Z, \mathbb{R}^P)$. We interpret (1.1) in the mild sense: the solution of (1.1) is given by

$$z(t) = S(t)z_0 + \int_0^t S(t-s)Bu(s)\, ds. \tag{1.3}$$

In this paper, we are concerned with a finite-dimensional compensator design for the system (1.1) and (1.2); i.e., we consider a finite-dimensional compensator of the form

$$\begin{aligned} \frac{d}{dt}w(t) &= (A_c - B_cK_c)w(t) + G_c(y - C_cw), \\ u(t) &= -K_cw(t), \end{aligned} \tag{1.4}$$

where $w(t) \in W = \mathbb{R}^{n_c}$ and A_c, B_c, C_c, K_c, G_c are the appropriate matrices. Hence, we obtain the closed-loop system on $Z \times W$

$$\frac{d}{dt}\begin{bmatrix} z(t) \\ w(t) \end{bmatrix} = H_c \begin{bmatrix} z(t) \\ w(t) \end{bmatrix}, \tag{1.5}$$

where

$$H_c = \begin{bmatrix} A & -BK_c \\ G_cC & F_c \end{bmatrix}$$

with $F_c = A_c - B_cK_c - G_cC_c$ and $\operatorname{dom}(H_c) = \operatorname{dom}(A) \times W$. This operator generates a strongly continuous semigroup on $Z \times W$ since it is a bounded perturbation of $H_0 = \left[\begin{smallmatrix} A & 0 \\ 0 & 0 \end{smallmatrix}\right]$.

* Received by the editors November 11, 1987; accepted for publication (in revised form) January 8, 1990.

† Center for Control Sciences, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This research was supported in part by Air Force Office of Scientific Research grants AFOSR-84-0398 and AFOSR-85-0303, and National Aeronautics and Space Administration grant NAG-1-517.

An aim of this paper is to establish the existence result of finite-dimensional compensators (1.4) for (1.1) and (1.2) (in § 2) such that the closed-loop operator H_c generates a uniformly exponentially stable semigroup on $Z \times W$. The semigroup $S(t)$ is said to be uniformly exponentially stable if the growth constant ω_0:

$$\omega_0 = \lim \frac{1}{t} \log \|S(t)\| \quad \text{as } t \to \infty,$$

is negative. A semigroup will simply be called stable if it is uniformly exponentially stable. Existence results have been established in [4], [16], and [19] provided that A has a complete set of generalized eigenvectors, and their construction procedure is based on the eigenvectors of either A [4], [16] or the closed-loop operator $A - BK$ [19]. In this paper, we will take a different approach which uses the finite-dimensional approximations of the system (1.1) and (1.2) via Galerkin-type approximation [1], [7]. The completeness assumption of eigenvectors of A can then be removed and will be replaced by the conditions on approximations. For example, we are able to prove the existence of a finite-dimensional compensator without the completeness assumption of eigenvectors for the system described by hereditary differential equations. Also, we will discuss the construction procedure that uses the approximating solutions of linear quadratic regulator and optimal observer, and a reduction of dimension n_c in (1.4) of compensators via the Schumacher approach [19] in § 3. In § 4, the general result is then applied to parabolic systems and hereditary differential systems. In [25], Bernstein and Hyland derived the optimal projection equation for finite-dimensional, fixed-order dynamic compensations for the system (1.1) and (1.2). The reduction procedure described in Theorem 3.5 is similar to the optimal projection schemes in [25]. Our procedure is not optimal in the sense of [25]. However, Theorem 3.5 provides a sufficient condition for the existence of an optimal fixed-order compensator and our approximation framework can be used to analyze convergence of fixed-order optimal compensator designs for approximate finite-dimensional systems to one for the original system (1.1) and (1.2). The approach used in this paper can be extended to a class of problems in which the input and output operators are unbounded employing the ideas in [5] and [15]. These regards as well as numerical examples will be reported elsewhere.

The approach we will take is based on the following facts. The pair (A, B) is said to be stabilizable if there exists an operator $K \in L(Z, \mathbb{R}^m)$ such that $A - BK$ generates a stable semigroup, and the pair (A, C) is detectable if there exists an operator $G \in L(\mathbb{R}^p, Z)$ such that $A - GC$ generates a stable semigroup on Z. If (A, B) is stabilizable and (A, C) is detectable, then there exists an infinite-dimensional compensator for (1.1) and (1.2); i.e., $w \in Z$ satisfies

$$\begin{aligned} &\frac{d}{dt} w(t) = (A - BK)w(t) + G(y - Cw(t)), \\ &u(t) = -Kw(t), \end{aligned} \tag{1.6}$$

which leads to a stable closed-loop system H. In fact, if we define the function $e = z - w$, then (1.1), (1.2), and (1.6) can be written as

$$\begin{aligned} &\frac{d}{dt} e = (A - GC)e, \\ &\frac{d}{dt} z = (A - BK)z + BKe. \end{aligned} \tag{1.7}$$

Thus, the stabilizability of (A, B) and the detectability of (A, C) imply the existence of infinite-dimensional compensators and, moreover, $\sigma(H) = \sigma(A - BK) \cup \sigma(A - GC)$. Next, we can construct a stabilizing feedback gain operator K by the solution of the linear quadratic regulator problem; consider the problem of minimizing the cost functional

$$J(u) = \int_0^\infty (\langle Qz(t), z(t)\rangle + |u(t)|^2)\, dt \tag{1.8}$$

subject to (1.3) where Q is a nonnegative (definite), bounded, self-adjoint operator on Z. Suppose (A, B) is stabilizable and (A, Q) is detectable, then the optimal solution of (1.8) is given by the feedback form $u(t) = -\hat{K}z(t)$ where $\hat{K} = B^*\Pi$ and Π is the unique, nonnegative, self-adjoint solution of the algebraic Riccati equation

$$(A^*\Pi + \Pi A - \Pi BB^*\Pi + Q)z = 0 \quad \text{for all } z \in \operatorname{dom}(A), \tag{1.9}$$

and $A - B\hat{K}$ generates a stable semigroup on Z (e.g., see [24]).

The approximation of the solution Π to the Riccati equation (1.9) has been studied (e.g., see [3], [6], [10]), which leads to a sequence $\hat{K}^N$ of finite-dimensional operators such that $\hat{K}^N$ converges to $\hat{K}$ in norm. Let Z^N be a sequence of finite-dimensional subspaces of Z, and let P^N denote the orthogonal projection of Z onto Z^N. Consider the approximating system (A^N, B^N, Q^N) where $A^N : Z^N \to Z^N$ is continuous, $B^N = P^N B$ and $Q^N = P^N Q P^N$. Under appropriate conditions on the triple (A^N, B^N, Q^N), which will be stated in § 3, it is shown in [10] that the unique nonnegative solution Π^N of the approximating Riccati equation in Z^N

$$A^{N*}\Pi^N + \Pi^N A^N - \Pi^N B^N B^{N*}\Pi^N + Q^N = 0, \tag{1.10}$$

converges strongly to Π, and for some constants $M \geqq 1$ and $\omega > 0$

$$\| e^{(A^N - B^N B^{N*}\Pi)t} P^N \| \leqq M e^{-\omega t}, \qquad t \geqq 0.$$

Since B is of finite rank, $\hat{K}^N = B^{N*}\Pi^N$ converges to $\hat{K}$ in norm. Similarly, we can apply the same procedure as above to the dual problem in order to obtain a convergent sequence $\hat{G}^N$ to $\hat{G}$ where the so-called Kalman filter gain $\hat{G}$ is given by $\hat{G} = \Sigma C^*$ and the self-adjoint operator Σ on Z satisfies

$$(A\Sigma + \Sigma A^* - \Sigma C^* C\Sigma + V)x = 0 \quad \text{for all } x \in \operatorname{dom}(A^*). \tag{1.11}$$

If (A, C) is detectable, V is a bounded, nonnegative, self-adjoint operator on Z, and (A, V) is stabilizable, (1.11) admits a unique nonnegative solution and $A - \hat{G}C$ generates a stable semigroup on Z. Let $C^N = CP^N$ and $V^N = P^N V P^N$. Then, $\hat{G}^N$ is given by $\hat{G}^N = \Sigma^N C^{N*}$ where Σ^N satisfies

$$A^N\Sigma^N + \Sigma^N A^{N*} - \Sigma^N C^{N*} C^N \Sigma^N + V^N = 0. \tag{1.12}$$

Let us denote by A_c^N, B_c^N, C_c^N, K_c^N, G_c^N, the matrix representation of A^N, B^N, C^N, K^N, G^N, respectively. We obtain a design of compensator (1.4) where $W = \mathbb{R}^{k_N}$ with $k_N = \dim(Z^N)$. Here, we may argue that for N sufficiently large, the closed-loop operator

$$H^N = \begin{bmatrix} A & -BK^N \\ G^N C & F^N \end{bmatrix} \quad \text{with } F^N = A^N - B^N K^N - G^N C^N \tag{1.13}$$

generates a stable semigroup on $Z \times Z^N$. We will give a sufficient condition for this claim in §§ 2 and 3.

2. Existence result. Consider the following hypotheses:

(H1) For every $z \in Z$, $e^{A^N t} P^N z$ converges strongly to $S(t)z$ where the convergence is uniform on bounded t-intervals.

(H2) There exists a sequence $K^N \in L(Z^N, \mathbb{R}^m)$ and $K \in L(Z, \mathbb{R}^m)$ such that $\|K^N - K\| \to 0$ as $N \to \infty$,

$$\|e^{(A^N - B^N K^N)t} P^N\| \leqq M_1 e^{-\omega_1 t}, \qquad t \geqq 0$$

for $M_1 \geqq 1$ and $\omega_1 > 0$, independent of N, and $A - BK$ generates a stable semigroup on Z with the growth constant $-\omega_3$.

(H3) There exists a sequence $G^N \in L(\mathbb{R}^P, Z^N)$ such that $\sup \|G^N\| < \infty$ and for $M_2 \geqq 1$ and $\omega_2 > 0$,

$$\|e^{(A^N - G^N C^N)t} P^N\| \leqq M_2 e^{-\omega_2 t}, \qquad t \geqq 0.$$

(H4) For any choice of the matrices A_c, B_c, C_c, K_c, and G_c in (1.4), H_c satisfies the spectrum-determined growth assumption [23]; the growth constant of the semigroup generated by H_c equals $\sup\{\operatorname{Re} \lambda : \lambda \in \sigma(H_c)\}$.

Since $Z^N \subset Z$, $P^N \phi = \phi$ for all $\phi \in Z^N$. Thus, K^N can be extended to all elements ϕ in Z by $K^N \phi = K^N P^N \phi$. Throughout the paper K^N will denote such an extension (i.e., $K^N = K^N P^N$).

Note that (H1) implies that P^N converges strongly to I, so that $\|B^N - B\|$ and $\|C^N - C\|$ converge to zero as $N \to \infty$. The following is the main result of the paper.

THEOREM 2.1. *Assume a family* (Z^N, A^N, B^N, C^N) *satisfies* (H1)-(H3) *with some* $M_1, M_2 \geqq 1$ *and* $\omega_1, \omega_2, \omega_3 > 0$. *Then for any* $\delta > 0$ *there exists an integer* N_δ *such that for* $N \geqq N_\delta$, *if* $\operatorname{Re} \lambda \geqq -\min(\omega_1, \omega_2, \omega_3) + \delta$, *then* $\lambda \in \rho(H^N)$. *Moreover, if* (H4) *is satisfied, then for* N *sufficiently large* H^N *generates a stable semigroup.*

Proof. First, we will show that for every $z \in Z$

$$e^{(A^N - B^N K^N)t} P^N z \to T(t)z,$$

uniformly on bounded t-intervals, where $T(t)$ is the semigroup generated by $A - BK$. By Theorem 4.4 [14, Chap. 3] it suffices to show that for some $\lambda > 0$

$$(\lambda I - (A^N - E^N))^{-1} P^N \to (\lambda I - (A - E))^{-1} \quad \text{(strongly)}, \tag{2.1}$$

where $E = BK$ and $E^N = B^N K^N$. Note that

$$(\lambda I - (A^N - E^N))^{-1} P^N - (\lambda I - A^N)^{-1} P^N = -(\lambda I - (A^N - E^N))^{-1} E^N (\lambda I - A^N)^{-1} P^N$$

and similarly

$$(\lambda I - (A - E))^{-1} - (\lambda I - A)^{-1} = -(\lambda I - (A - E))^{-1} E (\lambda I - A)^{-1}.$$

Thus,

$$\begin{aligned}
&(\lambda I - (A^N - E^N))^{-1} P^N - (\lambda I - (A - E))^{-1} \\
&\quad = -(\lambda I - (A^N - E^N))^{-1} E^N [(\lambda I - A^N)^{-1} P^N - (\lambda I - A)^{-1}] \\
&\qquad + (\lambda I - (A^N - E^N))^{-1} (P^N E - E^N)(\lambda I - A)^{-1} \\
&\qquad - [(\lambda I - (A^N - E^N))^{-1} P^N - (\lambda I - (A - E))^{-1}] E (\lambda I - A)^{-1} \\
&\qquad + (\lambda I - A^N)^{-1} P^N - (\lambda I - A)^{-1},
\end{aligned} \tag{2.2}$$

or

$$[(\lambda I-(A^N-E^N))^{-1}P^N-(\lambda I-(A-E))^{-1}](I+E(\lambda I-A)^{-1})$$
$$=[I-(\lambda I-(A^N-E^N))^{-1}E^N]((\lambda I-A^N)^{-1}P^N-(\lambda I-A)^{-1})$$
$$+(\lambda I(A^N-E^N))^{-1}(P^NE-E^N)(\lambda I-A)^{-1}.$$

Note that $I+E\ (\lambda I-A)^{-1}$ is continuously invertible; i.e., $(I+E(\lambda I-A)^{-1})^{-1}=(\lambda I-A)(\lambda I-(A-E))^{-1}$ where the right-hand side is continuous by the closed graph theorem. Hence (2.1) follows from the fact that $E^N\to E$, strongly and (H1), which is equivalent to $(\lambda I-A^N)^{-1}P^N\to(\lambda I-A)^{-1}$ by Theorem 4.4 of [14, Chap. 3].

Next we will show that for any $\delta>0$, there exists an integer N_δ such that if $N\geqq N_\delta$, then

$$\text{Re}\ \sigma(H^N)<-\min(\omega_1,\omega_2,\omega_3)+\delta. \tag{2.3}$$

For given $(f,g)\in Z\times Z^N$, consider the equation

$$\lambda\begin{bmatrix}\phi\\ \psi\end{bmatrix}-H^N\begin{bmatrix}\phi\\ \psi\end{bmatrix}=\begin{bmatrix}f\\ g\end{bmatrix}$$

for $(\phi,\psi)\in Z\times Z^N$, that is,

$$\lambda\phi-[(A-BK^N)\phi+BK^N(\phi-\psi)]=f, \tag{2.4}$$

$$\lambda\psi-[(A^N-B^NK^N)\psi+G^N(C\phi-C^N\psi)]=g, \tag{2.5}$$

where we used $K^N=K^NP^N$.

First note that since $\|K^N-K\|\to 0$ as $N\to\infty$, there exists an integer N_1 such that if $N\geqq N_1$ and $\text{Re}\ \lambda\geqq-\omega_3+\delta/2$, then $(\lambda I-(A-BK^N))^{-1}\in L(Z)$. In fact, by the variation of constants formula

$$T^N(t)=T(t)+\int_0^t T(t-s)B(K-K^N)T^N(s)\,ds,$$

where $T^N(t)$ is the semigroup generated by $A-BK^N$, and hence for $M_3\geqq 1$ and $\varepsilon>0$

$$\|T^N(t)\|\leqq M_3\,e^{(-\omega_3+\varepsilon)t}+\int_0^t M_3\,e^{(-\omega_3+\varepsilon)(t-s)}\|B\|\,\|K-K^N\|\,\|T^N(s)\|\,ds.$$

It then follows from Gronwall's lemma that

$$\|T^N(t)\|\leqq M_3\,e^{(-\omega_3+\varepsilon+M_3\|B\|\,\|K-K^N\|)t},\qquad t\geqq 0.$$

Now, for $N\geqq N_1$ and $\text{Re}\ \lambda\geqq-\omega_3+\delta/2$, from (2.4)

$$\phi=(\lambda I-(A-BK^N))^{-1}[BK^N(\phi-\psi)+f]. \tag{2.6}$$

From (2.5), if $\text{Re}\ \lambda>-\omega_1$

$$\psi=(\lambda I-(A^N-B^NK^N))^{-1}[G^NC(\phi-\psi)+g].$$

Thus, if $\text{Re}\ \lambda\geqq-\min(\omega_1,\omega_3)+\delta/2:=-\omega$ and $N\geqq N_1$

$$(\phi-\psi)=(\lambda I-(A-BK^N))^{-1}[BK^N(\phi-\psi)+f]$$
$$-(\lambda I-(A^N-B^NK^N))^{-1}[G^NC(\phi-\psi)+g],$$

or

$$
\begin{aligned}
&(\lambda I-(A^N-G^NC^N))(\phi-\psi)\\
&\quad=(\lambda I-(A^N-B^NK^N))[(\lambda I-(A-BK^N))^{-1}BK^N\\
&\qquad-(\lambda I-(A^N-B^NK^N))^{-1}B^NK^N](\phi-\psi)\\
&\qquad+(\lambda I-(A^N-B^NK^N))[(\lambda I-(A-BK^N))^{-1}f\\
&\qquad-(\lambda I-(A^N-B^NK^N))^{-1}P^Nf]\\
&\qquad+P^Nf-g+G^N(C^N-C)(\phi-\psi).
\end{aligned}
$$

Hence, we have for $\operatorname{Re}\lambda \geqslant -\omega$ and $N \geqq N_1$,

$$
\begin{aligned}
&(\phi-\psi)-\delta^NBK^N(\phi-\psi)+(\lambda I-(A^N-G^NC^N))^{-1}G^N(C-C^N)(\phi-\psi)\\
&\quad=(\lambda I-(A^N-G^NC^N))^{-1}(P^Nf-g)+\delta^Nf
\end{aligned}
\tag{2.7}
$$

where

$$
\begin{aligned}
\delta^N=&[I+(\lambda I-(A^N-G^NC^N))^{-1}(B^NK^N-G^NC^N)]\\
&\times((\lambda I-(A-BK^N))^{-1}-(\lambda I-(A^N-B^NK^N))^{-1}P^N).
\end{aligned}
\tag{2.8}
$$

Note that for $\operatorname{Re}\lambda > -\omega_2$

$$
\|(\lambda I-(A^N-G^NC^N))^{-1}P^N\| \leqq \frac{M_2}{\operatorname{Re}\lambda+\omega_2}.
$$

Thus,

$$
\|I+(\lambda I-(A^N-G^NC^N))^{-1}(G^NC^N+B^NK^N)\| \leqq \alpha
$$

for some constant α independent of N and $\operatorname{Re}\lambda \geqq -\omega_2+\delta$. Now we will show that there exists an integer $N_2(\geqq N_1)$ such that if $N \geqq N_2$, then $\|\delta^NBK^N\| \leqq \frac{1}{3}$.

For $\operatorname{Re}\lambda \geqq -\omega$

$$
(\lambda I-(A-BK^N))^{-1}B=\int_0^\infty e^{-\lambda t}T^N(t)B\,dt, \quad \text{and}
$$

$$
(\lambda I-(A^N-B^NK^N))^{-1}P^NB=\int_0^\infty e^{-\lambda t}\,e^{(A^N-B^NK^N)t}P^NB\,dt.
$$

Hence, for any $\tau>0$

$$
\begin{aligned}
&\|(\lambda I-(A-BK^N))^{-1}B-(\lambda I-(A^N-B^NK^N))^{-1}B^N\|\\
&\quad\leqq\int_0^\tau e^{-\operatorname{Re}\lambda t}\|e^{(A^N-B^NK^N)t}P^NB-T^N(t)B\|\,dt\\
&\qquad+\left(M_3\frac{e^{-(\operatorname{Re}\lambda+\omega)\tau}}{\operatorname{Re}\lambda+\omega}+M_1\frac{e^{-(\operatorname{Re}\lambda+\omega_1)\tau}}{\operatorname{Re}\lambda+\omega_1}\right)\|B\|.
\end{aligned}
\tag{2.9}
$$

Let us choose $\tau>0$ such that for $\operatorname{Re}\lambda > -\omega+\delta/2$

$$
\alpha\beta\left(M_1\frac{e^{-(\operatorname{Re}\lambda+\omega_1)\tau}}{\operatorname{Re}\lambda+\omega_1}+M_3\frac{e^{-(\operatorname{Re}\lambda+\omega)\tau}}{\operatorname{Re}\lambda+\omega}\right)\|B\| \leqq \frac{1}{6}
$$

where $\beta=\max\|K^N\|$. This is possible since $\operatorname{Re}\lambda+\omega$ and $\operatorname{Re}\lambda+\omega_1 \geqq \delta/2$. Next, the first term of (2.9) is bounded by

$$
\max(1, e^{-\operatorname{Re}\lambda\tau})\int_0^\tau \|e^{(A^N-B^NK^N)t}P^NB-T^N(t)B\|\,dt.
$$

Since $e^{(A^N-B^NK^N)t}P^NB$ and $T^N(t)B$ converge strongly to $T(t)B$, uniformly on $[0, \tau]$, and the range of B is finite ($=m$), there exists an integer N_2 such that if $N \geqq N_2$, then the first term of (2.9) is bounded by $1/6\alpha\beta$. It then follows that for $N \geqq N_2$ (depending only on δ), $\|\delta^N BK^N\| \leqq \frac{1}{3}$. Since $\|C^N - C\| \to 0$ as $N \to \infty$, for N sufficiently large, $\|(\lambda I - (A^N - G^NC^N))^{-1}G^N(C - C^N)\| \leqq \frac{1}{3}$ for $\operatorname{Re}\lambda \geqq -\omega_2 + \delta$, so we obtain that for any $\delta > 0$ there exists an integer N_δ such that if $N \geqq N_\delta$ and $\operatorname{Re}\lambda \geqq -\min(\omega_1, \omega_2, \omega_3) + \delta$, then

$$\|(I - \delta^N BK^N + (\lambda I - (A^N - G^NC^N))^{-1}G^N(C - C^N))^{-1}\| \leqq 3.$$

Therefore, it follows from (2.6) and (2.7) that if $\operatorname{Re}\lambda \geqq -\min(\omega_1, \omega_2, \omega_3) + \delta$ and $N \geqq N_\delta$, then $\lambda \in \rho(H^N)$, so that $\sup\{\operatorname{Re}\lambda : \lambda \in \sigma(H^N)\} \leqq -\min(\omega_1, \omega_2, \omega_3) + \delta$. Moreover, if (H4) is satisfied, then H^N generates a stable semigroup on $Z \times Z^N$. □

COROLLARY 2.2. *In addition to* (H1)-(H3) *we assume that* G^N *converges strongly to* G *and* $A - GC$ *generates a stable semigroup on* Z. *Then for every* $(f, g) \in Z \times Z$ *and* $\operatorname{Re}\lambda > -\min(\omega_1, \omega_2, \omega_3, \omega_4)$ *as* $N \to \infty$

$$(\lambda I - H^N)^{-1}\begin{pmatrix} f \\ P^N g \end{pmatrix} \to (\lambda I - H)^{-1}\begin{pmatrix} f \\ g \end{pmatrix} \quad \textit{strongly,}$$

where $-\omega_4$ *is the growth constant of the semigroup generated by* $A - GC$.

Proof. As in the proof of Theorem 2.1, we can show that

$$e^{(A^N - G^NC^N)t}P^N \to e^{(A-GC)t} \quad \text{strongly.}$$

Since for any $\varepsilon > 0$ there exists a constant $M_\varepsilon > 0$ such that $\|e^{(A-GC)t}\| \leqq M_\varepsilon e^{(-\omega_4+\varepsilon)t}$, it follows from Theorem 4.2 of [14, Chap. 3] that for $\operatorname{Re}\lambda > -\min(\omega_2, \omega_4)$

$$(\lambda I - (A^N - G^NC^N))^{-1} \to (\lambda I - (A - GC))^{-1} \quad \text{strongly.}$$

Recall that $\sigma(H) = \sigma(A - BK) \cup \sigma(A - GC)$. Thus, if $\operatorname{Re}\lambda < -\min(\omega_3, \omega_4)$, then $\lambda \in \rho(H)$ and using the same arguments as in the proof of Theorem 2.1 we obtain

$$(\lambda I - H)^{-1}\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} \phi \\ \psi \end{pmatrix}$$

where

$$\begin{aligned} \phi &= (\lambda I - (A - BK))^{-1}[BK(\phi + \psi) + f], \\ \phi - \psi &= (\lambda I - (A - GC))^{-1}(f - g). \end{aligned} \tag{2.10}$$

It follows from (2.9) that for $\operatorname{Re}\lambda > -\min(\omega_1, \omega_3)$ $\|\delta^N B\| \to 0$ as $N \to \infty$, where δ^N is defined by (2.8). Thus, the corollary follows from (2.6), (2.7), and (2.10).

3. Construction procedure. The following results are proved in Corollary 2.2 and Theorem 2.3 of [10].

THEOREM 3.1. *Assume the following*:

(A1) *For every* $z \in Z$, $e^{A^N t}P^N z$ *converges strongly to* $S(t)z$ *and* $e^{A^{N*}t}P^N z$ *converges strongly to* $S^*(t)z$ *where the convergence is uniform in* t *on bounded intervals.*

(A2) (A^N, B^N) *is uniformly stabilizable, i.e., there exists a sequence* $K^N \in L(Z^N, \mathbb{R}^m)$ *such that* $\sup \|K^N\| < \infty$ *and for* $M_1 \geqq 1$ *and* $\omega_1 > 0$

$$\|e^{(A^N - B^NK^N)t}P^N\| \leqq M_1 e^{-\omega_1 t}, \qquad t \geqq 0;$$

(A^N, C^N) is uniformly detectable, i.e., there exists a sequence $G^N \in L(\mathbb{R}^N, Z^N)$ such that $\sup \|G^N\| < \infty$ *and for $M_2 \geqq 1$ and $\omega_2 > 0$*

$$\|e^{(A^N - G^N C^N)t} P^N\| \leqq M_2 e^{-\omega_2 t}, \qquad t \geqq 0.$$

(A3) *(A^N, Q^N) is uniformly detectable and (A^N, V^N) is uniformly stabilizable.*

Then, (1.10) *has the unique nonnegative solution Π^N and for $\hat{M}_1 \geqq 1$ and $\hat{\omega}_1 > 0$*

$$\|e^{(A^N - BB^{N*}\Pi^N)t} P^N\| \leqq \hat{M}_1 e^{-\hat{\omega}_1 t}, \qquad t \geqq 0.$$

Also, (1.12) *has the unique nonnegative solution Σ^N and for $\hat{M}_2 \geqq 1$ and $\hat{\omega}_2 > 0$*

$$\|e^{(A^N - \Sigma^N C^{N*} C^N)t} P^N\| \leqq \hat{M}_2 e^{-\hat{\omega}_2 t}, \qquad t \geqq 0.$$

Moreover, (A, B) is stabilizable and (A, C) is detectable. If (A, Q) is detectable, then (1.9) *has the unique nonnegative solution Π and $\Pi^N P^N$ converges strongly to Π. Also, if (A, V) is stabilizable, then* (1.11) *has the unique nonnegative solution Σ and $\Sigma^N P^N$ converges strongly to Σ.*

Remark 3.2. Obviously, if Q and V are uniformly positive definite, then (A3) is satisfied, (A, Q) is detectable, and (A, V) is stabilizable.

The following theorem follows from Theorems 2.1 and 3.1.

THEOREM 3.3. *Assume that* (A1)-(A3) *are satisfied and that (A, Q) is detectable and (A, V) is stabilizable. Let $\hat{K} = B^*\Pi$ and $\hat{G} = \Sigma C^*$ and for each N let $\hat{K}^N = B^{N*}\Pi^N$ and $\hat{G}^N = \Sigma^N C^{N*}$. Then $\|\hat{K}^N - \hat{K}\|$, $\|\hat{G}^N - \hat{G}\|$ converge to zero as $N \to \infty$.*

Define an operator $\hat{H}^N$ defined on $Z \times Z^N$ by

$$\hat{H}^N = \begin{bmatrix} A & -B\hat{K}^N \\ \hat{G}^N C & \hat{F}^N \end{bmatrix} \quad \text{where } \hat{F}^N = A^N - B^N \hat{K}^N - \hat{G}^N C^N. \tag{3.1}$$

Then for any $\delta > 0$ there exists an integer N_δ such that for $N \geqq N_\delta$, $\{\lambda : \operatorname{Re} \lambda \geqq -\min(\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3) + \delta\}$ is contained in $\rho(\hat{H}^N)$. Here $-\hat{\omega}_3$ is the growth constant of the semigroup generated by $A - B\hat{K}$. Moreover, if (H4) *is satisfied, then $\hat{H}^N$ generates a stable semigroup on $Z \times Z^N$ for N sufficiently large.*

Next, we consider the reduction of dimension of compensators. In the remainder of this section, we assume the following:

(A4) For some $\lambda_0 > \omega_0$, $(\lambda_0 I + A)^{-1}$ is compact and

$$\|(\lambda_0 I - A^N)^{-1} P^N - (\lambda_0 I - A)^{-1}\| \to 0 \quad \text{as } N \to \infty$$

where A and $A^N \in G(M, \omega_0)$.

Then (A4) holds for every $\lambda > \omega_0$ [8, Thm. IV-2.25] and the spectrum of A consists entirely of isolated eigenvalues with finite multiplicities. It follows from Theorem 4.3 of [14, Chap. 3] that (A4) implies (A1).

LEMMA 3.4. *Assume* (A4) *is satisfied. Let E^N be a sequence in $L(Z^N)$ such that E^N converges strongly to $E \in L(Z)$. Then for some $\lambda > \omega_0$,*

$$\|(\lambda I - (A^N - E^N))^{-1} P^N - (\lambda I - (A - E))^{-1}\| \to 0 \quad \text{as } N \to \infty.$$

Proof. Since $\|E^N\| \to \|E\|$, for $\lambda > \omega_0 + \|E\| + \delta$, $\delta > 0$, $\lambda \in \rho(A^N - E^N)$. It has been shown in the proof of Theorem 2.1 that $(\lambda I - (A^N - E^N))^{-1} P^N$ converges strongly to $(\lambda I - (A - E))^{-1}$. Thus, $(\lambda I - (A^N - E^N))^{-1} P^N$ is uniformly bounded. It then follows from (2.2) that

$$\begin{aligned} &\|(\lambda I - (A^N - E^N))^{-1} P^N - (\lambda I - (A - E))^{-1}\| \\ &\quad \leqq \|I - (\lambda I - (A^N - E^N))^{-1} E^N\| \, \|(\lambda I - A^N)^{-1} P^N - (\lambda I - A)^{-1}\| \\ &\qquad + \|[(\lambda I - (A^N - E^N))^{-1} P^N - (\lambda I - (A - E))^{-1}] E (\lambda I - A)^{-1}\| \\ &\qquad + \|(\lambda I - (A^N - E^N))^{-1} P^N\| \, \|(E^N - E)(\lambda I - A)^{-1}\|. \end{aligned}$$

Since $(\lambda I - A)^{-1}$ and $E(\lambda I - A)^{-1}$ are compact, it follows that the second and third terms of the right-hand side of this inequality converge to zero. □

The approach of Schumacher in [19] can be used for reducing the dimension of compensators.

THEOREM 3.5. *Assume* (A2)-(A4) *are satisfied. Let* $\hat{K}$, $\hat{G}$, $\hat{K}^N$, *and* $\hat{G}^N$ *be given as in Theorem* 3.3. *Assume that* (i) *there exists a constant* $\delta_0 < 0$ *such that the halfplane* $\operatorname{Re}\lambda > \delta_0$ *contains only finitely many eigenvalues of* $A - BK$, *and* (ii) *there exists a constant* $r > 0$ *such that the eigenvalues of* $A - BK$ *and* $A^N - B^N K^N$ *with* $\operatorname{Re}\lambda > \delta_0$ *are in* $|\lambda| < r$.

For any δ, $\delta_0 < \delta < 0$, *let* V_δ *be the subspace of* $\operatorname{dom}(A)$ *spanned by the principal subspaces of* $A - B\hat{K}$ *for eigenvalues* λ *with* $\operatorname{Re}\lambda > \delta$ *and let* G_δ *be obtained by orthogonally projecting* $\hat{G}$ *onto* V_δ. *Similarly, let* V_δ^N *be the subspace of* Z^N *spanned by the principal subspaces of* $A^N - B^N\hat{K}^N$ *for eigenvalues* λ *with* $\operatorname{Re}\lambda > \delta$ *and* $\lambda \neq 0$, *and let* $\hat{G}_\delta^N$ *be obtained by orthogonally projecting* $\hat{G}^N$ *onto* V_δ^N.

Then for N *sufficiently large,* $\dim(V_\delta^N) = \dim(V_\delta) = q$ *and* $\|\hat{G}_\delta^N - \hat{G}_\delta\| \to 0$ *as* $N \to \infty$. *Moreover, if for some* $\delta A - \hat{G}_\delta C$ *generates a stable semigroup and* (H4) *holds, then for* N *sufficiently large the closed-loop operator*

$$\hat{H}_\delta^N = \begin{pmatrix} A & -B\hat{K}^N i^{N-1} \\ i^N \hat{G}_\delta^N C & i^N (A^N - B^N\hat{K}^N - \hat{G}_\delta^N C^N) i^{N-1} \end{pmatrix} \tag{3.2}$$

generates a stable semigroup on $Z \times \mathbb{R}^q$. *Here* V_δ^N *is identified with a Euclidean space* $\mathbb{R}^q$ *by the isomorphism* $i^N : V_\delta^N \to \mathbb{R}^q$.

Proof. Let Γ_0 be the boundary of the set $S = \{\operatorname{Re}\lambda \geq \delta_0 + \varepsilon\} \cap \{|\lambda| \leqq r\}$. Then we can choose $\varepsilon > 0$ so that $\Gamma_0 \subset \rho(A - B\hat{K})$ precisely encloses the eigenvalues of $A - B\hat{K}$ contained in $\operatorname{Re}\lambda > \delta_0$. Since $B^N\hat{K}^N \to B\hat{K}$ strongly by Theorem 3.1, it follows from Lemma 3.4 that for some $\lambda > \omega_0$

$$\|(\lambda I - (A^N - B^N\hat{K}^N))^{-1}P^N - (\lambda I - (A - B\hat{K}))^{-1}\| \to 0 \quad \text{as } N \to \infty. \tag{3.3}$$

This implies that (3.3) holds uniformly in λ on any compact subset of $\rho(A - B\hat{K})$. Thus, it follows from Theorem IV-3.16 of [8] that for N sufficiently large, $\Gamma_0 \subset \rho(A^N - B^N\hat{K}^N)$, the spectral projection

$$E_0^N = \frac{1}{2\pi i}\int_{\Gamma_0} (zI - (A^N - B^N\hat{K}^N))^{-1}P^N\,dz$$

exists, E_0^N converges to E_0 in norm, and $\dim \operatorname{range}(E_0^N) = \dim \operatorname{range}(E_0) < \infty$ where

$$E_0 = \frac{1}{2\pi i}\int_{\Gamma_0} (zI - (A - B\hat{K}))^{-1}\,dz.$$

These results now imply that for any δ, $\delta_0 < \delta < 0$, there exists a closed curve Γ_δ such that $\Gamma_\delta \subset \rho(A - B\hat{K})$ and $\rho(A^N - B^N\hat{K}^N)$ and it encloses precisely the eigenvalues of $A - BK$ and $A^N - B^N K^N$ with $\operatorname{Re}\lambda > \delta$ for N sufficiently large, and that if

$$E_\delta^N = \frac{1}{2\pi i}\int_{\Gamma_\delta} (zI - (A^N - B^N\hat{K}^N))^{-1}P^N\,dz, \quad \text{and}$$

$$E_\delta = \frac{1}{2\pi i}\int_{\Gamma_\delta} (zI - (A - B\hat{K}))^{-1}\,dz,$$

then E_δ^N converges to E_δ in norm and $\dim(V_\delta^N) = \dim(V_\delta)$ where $V_\delta^N = \operatorname{range} E_\delta^N$ and $V_\delta = \operatorname{range} E_\delta$. Thus, since $\|\hat{G}^N - \hat{G}\| \to 0$ as $N \to \infty$, $\|\hat{G}_\delta^N - \hat{G}_\delta\| \to 0$.

Let i be the isomorphism between V_δ and $W=\mathbb{R}^q$ and define

$$H_\delta=\begin{bmatrix} A & -B\hat{K}i^{-1} \\ iG_\delta C & i(A-B\hat{K}-\hat{G}_\delta C)i^{-1}\end{bmatrix}. \tag{3.4}$$

Then, since $\|\hat{G}_\delta^N-\hat{G}_\delta\|$, $\|\hat{K}^N-\hat{K}\|$, and $\|i^{N*}-i^*\|\to 0$ as $N\to\infty$, it is easy to show that

$$\|H_\delta^N-H_\delta\|\to 0 \quad \text{as } N\to\infty. \tag{3.5}$$

It follows from the proof of Theorem 4.2 of [19] that $\sigma(H_\delta)=\sigma(A-\hat{G}_\delta C)\cup\{$the eigenvalues of $A-B\hat{K}$ contained in $\operatorname{Re}\lambda>\delta\}$. Thus, if $A-\hat{G}_\delta C$ generates a stable semigroup on Z, then from the hypothesis (H4) H_δ generates a stable semigroup. The last statement of the theorem follows from (3.5), the variation of constant formula and Gronwall's lemma. □

4. Examples. In this section, we apply the general results in §§ 2 and 3 to two specific examples: the systems described by parabolic equations, and hereditary differential equations.

4.1. Parabolic systems. Let V and H be Hilbert spaces with V dense in H and assume the injection $i: V\to H$ is compact. Consider a sesquilinear form $\sigma: V\times V\to R$ such that

$$\sigma(u,v)\leqq C\|u\|_V\|v\|_V \quad \text{for } u,v\in V, \tag{4.1}$$

$$\sigma(u,u)\geqq \omega\|u\|_V^2-\rho\|u\|_H^2 \quad \text{for } u\in V \tag{4.2}$$

for $\omega>0$. It then follows from [22] that there exists an operator $A\in L(V,V^*)$ such that

$$\sigma(u,v)=\langle -Au,u\rangle \quad \text{for } u,v\in V \tag{4.3}$$

where $V\subset H=H^*\subset V^*$ and H is the pivoting space, and that A on H with

$$\operatorname{dom}(A)=\{x\in H: Ax\in H\}\text{: dense in } V \tag{4.4}$$

generates an analytic semigroup on H and V^*. Let Z^N be a sequence of finite-dimensional subspaces of V and let P^N be the orthogonal projection of H onto Z^N. Define $A^N: Z^N\to Z^N$ by

$$\langle -A^Nz,x\rangle=\sigma(z,x) \quad \text{for all } z,x\in Z^N. \tag{4.5}$$

For given $B\in L(\mathbb{R}^m,H)$ and $C\in L(H,\mathbb{R}^P)$ we define $B^N=P^NB$ and $C^N=CP^N$. We assume the approximation conditions:

$$\text{(C1)}\qquad \inf_{x\in Z^N}\|(\rho I-A)^{-1}z-x\|_V\leqq\varepsilon_1(N)\|z\|_H,$$

$$\text{(C2)}\qquad \inf_{x\in Z^N}\|(\rho I-A^*)^{-1}z-x\|_V\leqq\varepsilon_2(N)\|z\|_H,$$

where $\varepsilon_1(N)$, $\varepsilon_2(N)\to 0$ as $N\to\infty$. By the Nitsche technique (e.g., see [21]), we have

$$\|(\rho I-A^N)^{-1}P^N-(\rho I-A)^{-1}\|\leqq\frac{C^2}{\omega}\varepsilon_1(N)\varepsilon_2(N)\to 0. \tag{4.6}$$

Condition (4.2) shows that $(\rho I-A)^{-1}\in L(V^*,V)$. Since the injection i is compact, it follows that $(\rho I-A)^{-1}: H\to H$ is compact. Thus, in this case (A4) is satisfied. It follows from Theorem 6.A of [20] that

$$\|(\lambda I-A)^{-1}\|\leqq\frac{M}{|\lambda-\rho|} \quad \text{for } \lambda-\rho\in\left\{z\in\mathbb{C}: |\arg z\|<\frac{\pi}{2}+\theta_0\right\},$$

where $\theta_0 = \tan^{-1}(\omega/2C)$ and

$$S(t) = \frac{1}{2\pi i}\int_\Gamma e^{\lambda t}(\lambda I - A)^{-1}\,d\lambda, \qquad t>0,$$

where Γ is the path consisting of the two rays $|\arg(z-\rho)| = \pi/2+\theta$, $|z-\rho| \geqq 1$, $0<\theta<\theta_0$, and the semicircle $\{z-\rho = e^{it}: |t| \leqq \theta+\pi/2\}$ oriented so that $\operatorname{Im}\lambda$ increases along Γ. Suppose (A, B) is stabilizable: there exists an operator $K \in L(H, \mathbb{R}^m)$ such that $A-BK$ generates a stable semigroup on H. Thus, $\lambda_B = \sup\{\operatorname{Re}\lambda: \lambda \in (A-BK)\} < 0$. Define the sesquilinear form $\sigma_B(u, v)$ by

$$\sigma_B(u, v) = \sigma(u, v) + \langle BKu, v\rangle_H \quad \text{for } u, v \in V.$$

Then $\sigma_B(u, v) \geqq \omega\|u\|_V^2 - \rho_B|u|_H^2$ where $\rho_B = \rho + \|B\|\,\|K\|$, and $\sigma_B(u, v) = \langle -(A-BK)u, v\rangle$ for $u, v \in V$. If $T^N(t)$, $t \geqq 0$ is the semigroup generated by $A^N - B^N K^N$, with $K^N = KP^N$, then

$$T^N(t) = \frac{1}{2\pi i}\int_{\Gamma_B} e^{\lambda t}(\lambda I - (A^N - B^N K^N))^{-1}\,d\lambda$$

where the path $\Gamma_B = \{z \in \mathbb{C}: z = y + \|B\|\,\|K\|, y \in \Gamma\}$. From Lemma 3.4, we have

$$\|(\lambda I - (A^N - B^N K^N))^{-1}P^N - (\lambda I - A - BK)^{-1}\| \to 0 \quad \text{as } N \to \infty.$$

Thus, it again follows from Theorem IV-3.16 of [8] that we can shift the path Γ_B without changing the value of the integral to the path Γ'_B where for $\lambda_B < \lambda_0 < 0$ $\Gamma'_B = \{\Gamma \cap \{\operatorname{Re}\lambda \leqslant \lambda_0\}\} \cup \{\lambda \in \mathbb{C}: \operatorname{Re}\lambda = \lambda_0 \text{ and } |\operatorname{Im}\lambda| \leqq (\rho_B - \lambda_0)\tan\theta\}$, for N sufficiently large. From this, we can show that for N sufficiently large there exist constants $M_1 \geqq 1$ and $\lambda_B < -\omega_1 < 0$ such that

$$\|T^N(t)P^N\| \leqq M_1 e^{-\omega_1 t}, \qquad t \geqq 0.$$

Similarly, if (A, C) is detectable: $A-GC$ generates a stable semigroup for some $G \in L(\mathbb{R}^P, H)$, then for N sufficiently large

$$\|e^{(A^N - G^N C^N)t}P^N\| \leqq M_2 e^{-\omega_2 t}, \qquad t \geqq 0$$

for some constants $M_2 \geqq 1$ and $\omega_2 > 0$. Hence, (A2) is satisfied. Since $H_0 = \left[\begin{smallmatrix} A & 0 \\ 0 & 0 \end{smallmatrix}\right]$ generates an analytic semigroup on $Z \times W$ and H_c is a bounded perturbation of H_0, it follows from [23] that the hypothesis (H4) is satisfied. Therefore, Theorems 2.1, 3.3, and 3.5 are applied to the problem described above. Moreover, the strong version of Theorem 3.5 holds.

THEOREM 4.1. *Suppose* (A, B) *is stabilizable,* (A, C) *is detectable, and* (A3) *is satisfied. Let us adopt the same notation as in Theorem* 3.5. *If for some* $\delta < 0$ $\sup\{\operatorname{Re}\lambda: \lambda \in \sigma(A^N - \hat{G}_\delta^N C^N)\} < 0$ *for* N *sufficiently large, then* $\hat{H}_\delta^N$ *defined by* (3.2) *generates a stable semigroup.*

Proof. First we note that all the eigenvalues of $A-B\hat{K}$ and $A^N - B^N\hat{K}^N$ are contained in the sector $\{z: |\arg(z-\rho_1)| > \pi/2+\theta\}$ where $\rho_1 = \rho + \|B\| \sup\|\hat{K}^N\|$ and $0<\theta<\theta_0$ and that $A-B\hat{K}$ has a compact resolvent. So, for any $\delta_0 < 0$, the halfplane $\operatorname{Re}\lambda > \delta_0$ contains only finitely many eigenvalues of $A-B\hat{K}$. It then follows from Theorem 3.5 that for any $\delta < 0$, $\dim(V_\delta^N) = \dim(V_\delta)$ and $\|\hat{G}_\delta^N - \hat{G}_\delta\| \to 0$ as $N \to \infty$. Here,

$$\langle -(A - \hat{G}_\delta C)u, v\rangle = \sigma(u, v) + \langle \hat{G}_\delta Cu, v\rangle \quad \text{for } u, v \in V,$$
$$\langle -(A^N - \hat{G}_\delta^N C)z, x\rangle = \sigma(z, x) + \langle \hat{G}_\delta^N Cz, x\rangle \quad \text{for } z, x \in Z^N.$$

Thus, all the eigenvalues of $A-\hat{G}_\delta C$ and $A^N - \hat{G}_\delta^N C^N$ are contained in the sector $\{z: |\arg(z-\rho_2)| > \pi/2+\theta\}$ where $\rho_2 = \rho + \|C\| \sup\|\hat{G}^N\|$ and $A-\hat{G}_\delta C$ has a compact

resolvent. For $\gamma<0$ such that $\{\lambda: \operatorname{Re}\lambda=\gamma\}\subset\rho(A-\hat{G}_\delta C)$, let Γ be the closed path consisting of

$$\Gamma_\pm=\left\{\lambda=\gamma e^{\pm i\theta}+\rho_2, 0\leqq r\leqq\frac{\rho_2-\gamma}{\cos\theta}\right\}, \quad \text{and}$$

$$\Gamma_2=\{\operatorname{Re}\lambda=\gamma: |\operatorname{Im}\lambda|\leqq(\rho_2-\gamma)\tan\theta\}.$$

Then, Γ encloses only finitely many eigenvalues of $A-\hat{G}_\delta C$. It follows from Lemma 3.4 that

$$\|(\lambda I-(A^N-\hat{G}_\delta^N C^N))^{-1}P^N-(\lambda I-(A-\hat{G}_\delta C))^{-1}\|\to 0 \quad \text{as } N\to\infty$$

for some $\lambda>\omega_0$, and hence the convergence is uniform on Γ. It thus follows from Theorem IV-3.16 in [8] that if the convergence is uniform

$$E^N=\frac{1}{2\pi_i}\int_\Gamma (zI-(A^N-\hat{G}_\delta^N C^N))^{-1}P^N\,dz, \quad \text{and}$$

$$E=\frac{1}{2\pi i}\int_\Gamma (zI-(A-\hat{G}_\delta C))^{-1}\,dz,$$

then $\Gamma\subset\rho(A^N-\hat{G}_\delta^N C^N)$, $\dim \text{range}\,(E^N)=\dim\text{range}\,(E)=q$, and $\|E^N-E\|\to 0$ as $N\to\infty$. Thus Γ contains q eigenvalues counting according to algebraic multiplicities. Furthermore, let μ be an eigenvalue of $A-\hat{G}_\delta C$ such that $\operatorname{Re}\lambda\leqq\operatorname{Re}\mu$ for all $\lambda\in\sigma(A-\hat{G}_\delta C)$ and let Γ' be a circle centered at μ with an arbitrary small radius; then Γ' contains at least one eigenvalue of $A^N-\hat{G}_\delta^N C^N$ for N sufficiently large. Hence if $\sup\{\operatorname{Re}\lambda:\lambda\in\sigma(A^N-\hat{G}_\delta^N C^N)\}<0$, then $\operatorname{Re}\mu<0$. This implies that $\sup\{\operatorname{Re}\lambda:\lambda\in\sigma(A-\hat{G}_\delta C)\}<0$. Since $A-\hat{G}_\delta C$ generates an analytic semigroup, it follows from [23] that it generates a stable semigroup. Therefore, the theorem is a consequence of Theorem 3.5.

4.2. Hereditary differential systems. Consider the hereditary differential control system:

$$\begin{aligned}&\frac{d}{dt}x(t)=\int_{-r}^{0}d\mu(\theta)x(t+\theta)+Bu(t),\\ &x(0)=\eta \quad\text{and}\quad x(\theta)=\phi(\theta), \quad -r\leqq\theta<0,\end{aligned}\tag{4.7}$$

with the observation that

$$y(t)=Cx(t),$$

where $x(t)\in\mathbb{R}^n$, $u(t)\in\mathbb{R}^m$, $y(t)\in\mathbb{R}^p$, and $\mu(\theta)$ is an $n\times n$ matrix-valued function of bounded variation which vanishes at $\theta=0$ and is left continuous on $[-r,0]$. B and C are $n\times m$ and $p\times n$ matrices. We will denote by Z, the product space $\mathbb{R}^n\times L_2(-r,0;\mathbb{R}^N)$ in this section. It is well known [2] that for $(\eta,\phi)\in Z$ and u locally square integrable, (4.7) admits a unique solution $x\in L_2(-r,T;\mathbb{R}^n)\cap H^1(0,T;\mathbb{R}^n)$ and (4.7) can be equivalently formulated as an evolution equation on Z

$$\begin{aligned}&\frac{d}{dt}z(t)=Az(t)+Bu(t), \qquad z(0)=(\eta,\phi),\\ &y(t)=Cz(t),\end{aligned}\tag{4.8}$$

where $z(t)=(x(t),x(t,\cdot))\in Z$, $t\geqq 0$, for $u\in\mathbb{R}^m$; $Bu=(Bu,0)\in Z$, and for $(\eta,\phi)\in Z$, $C(\eta,\phi)=C\eta$. The infinitesimal generator A of the semigroup $S(t)$ is defined by

$$\text{dom}\,(A)=\{(\eta,\phi)\in Z: \dot{\phi}\in L_2 \text{ and } \eta=\phi(0)\}\tag{4.9}$$

and for $(\phi(0), \phi) \in \operatorname{dom}(A)$

$$A(\phi(0), \phi) = \left(\int_{-r}^{0} d\mu(\theta)\phi(\theta), \dot{\phi} \right). \tag{4.10}$$

As in [2], [6], [13], and [18], we consider the averaging approximation of (4.8). Let Z^N be a sequence of subspaces of Z defined by

$$Z^N = \{(\eta, \phi) \in Z: \phi(\theta) = a_j \text{ on } (\tau_j^N, \tau_{j-1}^N), j = 1, \cdots, N\}$$

and let P^N be the corresponding orthogonal projection of Z onto Z^N, where $\tau_j^N = -j(r/N)$. Z^N can be identified with $R^{n(N+1)}$ by means of the embedding $j^N : \mathbb{R}^{n(N+1)} \to Z^N$ defined by

$$j^N a = \left(a_0, \sum_{k=1}^{N} a_k \chi_{(\tau_k^N, \tau_{k-1}^N)} \right) \quad \text{for } a = (a_0^T, \cdots, a_N^T)^T \in R^{n(N+1)},$$

where χ_I denotes the characteristic function of an interval I. On $\mathbb{R}^{n(N+1)}$, we consider the induced inner product:

$$\langle x, w \rangle_N = x^T Q^N w, \quad x, w \in \mathbb{R}^{n(N+1)},$$

where

$$Q^N = \begin{bmatrix} I & & & \\ & \frac{r}{N} I & & \\ & & \ddots & \\ & & & \frac{r}{N} I \end{bmatrix}.$$

Then the adjoint operator j^{N*} is given by

$$[j^{N*}(\eta, \phi)]_0 = \eta \quad \text{and} \quad [j^{N*}(\eta, \phi)]_k = \frac{N}{r} \int_{\tau_k^N}^{\tau_{k-1}^N} \phi(\theta)\, d\theta, \qquad 1 \leqq k \leqq N.$$

It is easy to show that

$$j^{N*} j^N = id \quad \text{and} \quad j^N j^{N*} = P^N.$$

On Z^N, we consider the approximation of A:

$$A^N = j^N (Q^N)^{-1} H^N j^{N*} \tag{4.11}$$

where

$$H^N = \begin{bmatrix} A_0^N & A_1^N & \cdots & A_N^N \\ I & -I & & \\ & \ddots & \ddots & \\ & & I & -I \end{bmatrix}$$

and

$$A_k^N = \lim_{\tau \uparrow \tau_k^N} \left[\mu\left(\tau + \frac{r}{N}\right) - \mu(\tau) \right], \qquad k = 0, 1, \cdots, N.$$

In this case, $B^N = B$ and $C^N = C$. Then the following results have been proved.

THEOREM 4.2 (1) [18, Cor. 4.5]. *$\{\mathrm{Re}\,\lambda > \mathrm{Var}\,\mu\} \subset \rho(A)$ and $\rho(A^N)$ and for some $\lambda > \mathrm{Var}\,\mu$*

$$\|(\lambda I - A)^{-1} - (\lambda I - A^N)^{-1} P^N\| \to 0 \quad \text{as } N \to \infty. \tag{4.12}$$

(2) [13, Lemma 3.3]. *For $\gamma > \mathrm{Var}\,\mu$ and $b > r$, let $a_0 = b^{-1} \log[(1+\sqrt{2})(b-r)^{-1}]$ and $a_1 = \gamma + b^{-1} \log \mathrm{Var}\,\mu$ and for $a > \max(a_0, a_1)$ define the set*

$$\Sigma_a = \{\lambda \in \mathbb{C}: |\mathrm{Im}\,\lambda| \geqq e^{(a - \mathrm{Re}\,\lambda)b} \text{ and } \mathrm{Re}\,\lambda \leqq \gamma\}.$$

Then for every $\lambda \in \Sigma_a$

$$\|(\lambda I - A^N)^{-1} P^N\| \quad \text{and} \quad \|(\lambda I - A)^{-1}\| \leqq M |\mathrm{Im}\,\lambda|, \tag{4.13}$$

where M is independent of N and λ and is a continuous function of a, b, and γ. If Γ is the boundary of Σ_a oriented so that $\mathrm{Im}\,\lambda$ increases along Γ, then

$$\begin{aligned} e^{A^N t} P^N &= \frac{1}{2\pi i} \int_\Gamma e^{\lambda t} (\lambda I - A^N)^{-1} P^N \, d\lambda, \quad \text{and} \\ S(t) &= \frac{1}{2\pi i} \int_\Gamma e^{\lambda t} (\lambda I - A)^{-1} \, d\lambda. \end{aligned} \tag{4.14}$$

Since $(\lambda I - A)^{-1}$ is compact for some $\lambda > \mathrm{Var}\,\mu$ (A4) is satisfied. The following lemmas show that (A2) is satisfied in this example.

LEMMA 4.3. *If (A, B) is stabilizable, then (A^N, B^N) is uniformly stabilizable for $N \geqq N_0$.*

Proof. For some $K \in L(Z, \mathbb{R}^m)$ given by

$$K(\eta, \phi) = K_0 \eta + \int_{-r}^0 K(\theta) \phi(\theta) \, d(\theta), \qquad (\eta, \theta) \in Z,$$

$A - BK$ generates a stable semigroup. Note that the operator $A - BK$ associates with the hereditary differential equation of the form (4.7); i.e., for $\phi \in H^1$

$$(A - BK)(\phi(0), \phi) = \left(\int_{-r}^0 d\tilde{\mu}(\theta) \phi(\theta), \dot{\phi} \right),$$

where $\tilde{\mu} = \mu - B(\int_{-r}^{\bullet} K(\xi)\, d\xi + K_0 \mu_{]0]})$. Let $K^N = KP^N$; then $\|K^N - K\| \to 0$ as $N \to \infty$ and using exactly the same arguments as in the previous section for the parabolic case, it follows from Theorem 4.2 that for $0 > \omega > \sup\{\mathrm{Re}\,\lambda: \lambda \in \sigma(A - BK)\}$ there exists an integer N_0 such that if $N \geqq N_0$, then for $t > 2b$

$$e^{(A^N - B^N K^N)t} P^N = \frac{1}{2\pi i} \int_{\tilde{\Gamma}} e^{\lambda t} (\lambda I - (A^N - B^N K^N))^{-1} P^N \, d\lambda, \tag{4.15}$$

where the path $\tilde{\Gamma}$ consists of

$$\begin{aligned} \tilde{\Gamma}_1 &= \{\lambda: |\mathrm{Im}\,\lambda| = e^{(a - \mathrm{Re}\,\lambda)b} \text{ and } \mathrm{Re}\,\lambda \leqq \omega\}, \\ \tilde{\Gamma}_2 &= \{\lambda: \mathrm{Re}\,\lambda = \omega \text{ and } |\mathrm{Im}\,\lambda| \leqq e^{(a-\omega)b}\}, \\ \tilde{\Gamma}_3 &= \text{the mirror image of } \tilde{\Gamma}_1, \end{aligned}$$

and

$$\begin{aligned} &\|(\lambda I - (A^N - B^N K^N))^{-1} P^N\| \leqq M |\mathrm{Im}\,\lambda| \quad \text{on } \tilde{\Gamma}_1 \cup \tilde{\Gamma}_3, \\ &\|(\lambda I - (A^N - B^N K^N))^{-1} P^N\| \leqq M \quad \text{on } \tilde{\Gamma}_2, \end{aligned} \tag{4.16}$$

for some positive constants a and M. On $\tilde{\Gamma}_1 \cup \tilde{\Gamma}_3$

$$|\operatorname{Im}\lambda||e^{\lambda t}| \leqq e^{(a-\operatorname{Re}\lambda)b}\, e^{(\operatorname{Re}\lambda)t} = |\operatorname{Im}\lambda|^{(b-t)/b}\, e^{at},$$

so

$$\int_{\tilde{\Gamma}_1 \cup \tilde{\Gamma}_3} |\operatorname{Im}\lambda||e^{\lambda t}|\, d|\lambda| \leqq C\, e^{at} \int_{\{|\tau| \geqq e^{(a-\omega)t}\}} |\tau|^{(b-t)/b}\, dt$$

$$= \frac{2Cb}{t-2b}\, e^{(a-\omega)(2b-t)}\, e^{at} = \frac{2Cb}{t-2b}\, e^{2b(a-\omega)}\, e^{\omega t},$$

where C is a positive constant independent of t. It thus follows from (4.15) and (4.16) that

$$\|e^{(A^N - B^N K^N)t} P^N\| \leqq \tilde{M}\, e^{\omega t}, \qquad t \geqq 3b$$

for some constant $\tilde{M}$. Since $\max_{0 \leqq t \leqq 3b} \|e^{(A^N - B^N K^N)t} P^N\|$ are uniformly bounded in $N \geqq N_0$, the proof is complete.

LEMMA 4.4. *If (A, C) is detectable, then for $N \geqq N_1(A^N, C^N)$ is uniformly detectable.*

Proof. First we note that without loss of generality we can assume that for some $GZ \subset \operatorname{dom}(A)$, $A - GC$ generates a stable semigroup. In fact [9], [15] if (A, C) is detectable, then the Riccati equation

$$(A\Sigma + \Sigma A^* - \Sigma C^* C \Sigma + I I^*)z = 0 \quad \text{for all } z \in \operatorname{dom}(A^*)$$

has a unique nonnegative solution Σ such that $\Sigma z \in \operatorname{dom}(A)$ for every $z \in Z$ and $A - \Sigma C^* C$ generates a stable semigroup, where $Ix = (x, 0) \in Z$ for $x \in \mathbb{R}^n$.

Let us define $G^N \in L(\mathbb{R}^P, Z^N)$ by

$$[G^N y]_0 = G(0)y \quad \text{and} \quad [G^N y]_k = G(\tau_k^N), \quad 1 \leqq k \leqq N,$$

where $Gy = (G(0)y, G(\cdot)y) \in Z$, $y \in \mathbb{R}^P$. Then since $G(\cdot) \in H^1(-r, 0)$, $\|G^N - G\| \to 0$ as $N \to \infty$, and since

$$A^N G^N y = \left(\sum_{j=0}^{N} A_j^N G(\tau_j^N),\ \sum_{k=1}^{N} (G(\tau_{k-1}^N) - G(\tau_k^N))\chi_{(\tau_k^N, \tau_{k-1}^N)} \right),$$

$\|A^N G^N\| \leqq \alpha$ for some positive constant α. By the variation of constants formula

$$e^{(A^N - G^N C^N)t} = e^{A^N t} - \int_0^t e^{A^N(t-s)} G^N C^N\, e^{(A^N - G^N C^N)s}\, ds. \tag{4.17}$$

From (4.14) and Theorem 4.7 of [14, Chap. 2],

$$\|A^N\, e^{A^N t_0} P^N\| \leqq M_1 \quad \text{for some constant } M_1, \tag{4.18}$$

where $t_0 = 4b$. It follows from (4.17) and (4.18) that there exists a uniform constant M_2 such that

$$\|(A^N - G^N C^N)\, e^{(A^N - G^N C^N)t_0} P^N\| \leqq M_2.$$

Thus, from Theorem 4.7 of [14, Chap. 2], for $t > 2t_0$

$$e^{(A^N - G^N C^N)t} P^N = \frac{1}{2\pi i} \int_\Gamma e^{\lambda t} (\lambda I - (A^N - G^N C^N))^{-1} P^N\, d\lambda, \tag{4.19}$$

where the integral path Γ is given by $\Gamma=\Gamma_1\cup\Gamma_2\cup\Gamma_3$,

$$\Gamma_1=\{\lambda:\ \mathrm{Im}\,\lambda=e^{(\alpha-\mathrm{Re}\,\lambda)t_0}\ \text{and}\ \mathrm{Re}\,\lambda\leqq\beta\},$$

$$\Gamma_2=\{\lambda:\ \mathrm{Re}\,\lambda=\beta\ \text{and}\ |\mathrm{Im}\,\lambda|\leqq e^{(\alpha-\beta)t_0}\},$$

$$\Gamma_3=\text{the mirror image of}\ \Gamma_1.$$

$\alpha=\log(2M_2)$ and $\beta=\gamma+\sup\|G^N\|\,\|G\|$. And, for all $\lambda\in\Sigma_\alpha$

$$\|(\lambda I-(A^N-G^NC^N))^{-1}P^N\|\leqq 2M_2(1+t_0)\,e^{\beta t_0}|\mathrm{Im}\,\lambda|. \tag{4.20}$$

Similarly, since $GCy\in\mathrm{dom}\,(A)$, $y\in\mathbb{R}^P$, $A-GC$ generates a differentiable semigroup for $t>2t_0$. It then follows from [23] that if $\omega_0=\sup\{\mathrm{Re}\,\lambda:\lambda\in\sigma(A-GC)\}$, then $\omega_0<0$. From Lemma 3.4 with $E^N=G^NC^N$ and Theorem IV-3.15 of [8], for $\omega_0<\omega<0$ there exists an integer N_1 such that if $N\geqq N_1$, then $(\lambda I-(A^N-G^NC^N))^{-1}P^N$ is analytic on the compact set $\Sigma_\alpha\cap\{\mathrm{Re}\,\lambda\geqq\omega\}$. By the Cauchy theorem, we can, without changing the value of integral (4.19), shift Γ to the path Γ' for which the constant β is replaced by ω everywhere above. Hence, the lemma follows from the same calculations in the proof of Lemma 4.3. □

Let us consider in (1.9) and (1.11)

$$Q(\eta,\phi)=(Q_0,0)\quad\text{and}\quad V(\eta,\phi)=(V_0\eta,0)\quad\text{for}\ (\eta,\phi)\in Z, \tag{4.21}$$

where Q_0 and V_0 are nonnegative symmetric matrices on $\mathbb{R}^n$.

Remark 4.5. If Q_0, V_0 are full rank in (4.20), then (A, Q) is detectable and (A, V) is stabilizable by the rank condition [17]. Then, in this case, it follows from Lemma 4.3 and 4.4 that (A^N, Q^N) is uniformly detectable and (A^N, V^N) is uniformly stabilizable.

Now, the following theorem is a consequence of Theorem 2.1.

THEOREM 4.6. *Suppose* (A, B) *is stabilizable and* (A, V) *is detectable; then there exists a finite-dimensional compensator of the form* (1.4) *such that the closed-loop operator* H_c (*see* (1.5)) *generates a stable semigroup.*

Next, we prove the corresponding result to Theorem 4.1 for the hereditary differential system. To this end, we need the following lemma which also provides a new convergence result on the averaging approximation.

LEMMA 4.7. *Suppose* $V(\eta,\phi)=(V_0\eta,0)$ *for* $(\eta,\phi)\in Z$, *and* (A, V) *is stabilizable. Then if* Σ^N *is the nonnegative solution of* (1.12), $\|A^N\Sigma^N\|$ *are uniformly bounded in N.*

Proof. It follows from Lemmas 4.3 and 4.4 that for some positive integer N_2, if $N\geqq N_2$, then (A^N, C^N) is uniformly detectable and (A^N, V^N) is uniformly stabilizable. Thus, from Theorem 3.1, (1.12) has the unique nonnegative solution Σ^N and for $M\geqq 1$ and $\omega>0$

$$\|e^{(A^N-\Sigma^NC^{N*}C^N)t}P^N\|\leqq M\,e^{-\omega t},\qquad t\geqq 0.$$

Hence from (1.12)

$$\Sigma^N=\int_0^t e^{A^N(t-s)}V^N\,e^{(A^N-\Sigma^NC^{N*}C^N)*(t-s)}\,ds+e^{A^Nt}\Sigma^N\,e^{(A^N-\Sigma^NC^{N*}C^N)*t}. \tag{4.22}$$

We will show that

$$\left\|A^N\int_0^t e^{A^N(t-s)}If(s)\,ds\right\|\leqq C(t)\|f\|_{L^2(0,t)} \tag{4.23}$$

where $Ix=(x,0)\in Z$, $x\in\mathbb{R}^n$, and $C(t)$ is a nondecreasing function of $t\geqq 0$. For $t\geqq 0$

let $z^N(t) = \int_0^t e^{A^N(t-s)} If(s)\, ds$. Then $z^N(t)$ satisfies

$$\frac{d}{dt} z^N(t) = A^N z^N(t) + If(t), \qquad z^N(0) = 0.$$

Let us consider the operator A_0 defined on Z defined by

$$\begin{aligned} &\text{dom } (A_0) = \{(\eta, \phi) \in Z: \phi \in L^2 \text{ and } \phi(0) = \eta\}, \\ &A_0(\phi(0), \phi) = (\phi(0), \dot{\phi}) \quad \text{for } \phi \in H^1. \end{aligned} \tag{4.24}$$

From (4.11), $z^N(t) = j^N a^N(t)$ where $a^N \in R^{n(N+1)}$ satisfies

$$\frac{d}{dt} a_0 = \sum_{j=0}^{N} A_j^N a_j + f(t),$$

$$\frac{r}{N} \frac{d}{dt} a_k = a_{k-1} - a_k, \qquad 1 \leqq k \leqq N.$$

Thus, if we define

$$E(z^N) = \|A_0^N z^N\|_z^2 = |a_0|^2 + \sum_{k=1}^{N} \frac{N}{r} |a_{k-1} - a_k|^2, \tag{4.25}$$

then

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} E(z^N(t)) &= \langle \dot{a}_0, a_0 \rangle + \frac{N}{r} \sum_{k=1}^{N} \langle \dot{a}_{k-1} - \dot{a}_k, a_{k-1} - a_k \rangle \\ &= \langle \dot{a}_0, a_0 \rangle - \sum_{k=1}^{N} \langle \dot{a}_{k-1} - \dot{a}_k, \dot{a}_k \rangle \\ &= \langle \dot{a}_0, a_0 \rangle + \frac{1}{2} |\dot{a}_0|^2 - \frac{1}{2} \sum_{k=1}^{N} |\dot{a}_k - \dot{a}_{k-1}|^2 - \frac{1}{2} |\dot{a}_N|^2 \\ &\leqq \frac{1}{2} |a_0|^2 + |\dot{a}_0|^2. \end{aligned} \tag{4.26}$$

Note that for $1 \leqq k \leqq N$

$$\begin{aligned} |a_k|^2 &= \left| a_0 + \sum_{j=0}^{k} (a_j - a_{j-1}) \right|^2 \leqq 2|a_0|^2 + 2 \left| \sum_{j=1}^{k} (a_j - a_{j-1}) \right|^2 \\ &\leqq 2|a_0|^2 + 2k \sum_{j=1}^{k} |a_j - a_{j-1}|^2 \\ &\leqq 2 \max(1, r) E(z^N). \end{aligned}$$

Thus,

$$\begin{aligned} |\dot{a}_0| &\leqq \sum_{j=0}^{N} A_j^N |a_j| + |f| \\ &\leqq \sqrt{2 \max(1, r)}\, \text{Var } \mu + |f|. \end{aligned}$$

From (4.24)

$$\frac{1}{2} \frac{d}{dt} E(z^N(t)) \leqq \alpha E^N(z^N(t)) + 2|f|^2,$$

where $\alpha = \frac{1}{2} + 4 \max(1, r)(\mathrm{Var}\, \mu)^2$, so that by Gronwall's lemma

$$E(z^N(t)) \leqq e^{2\alpha t}\left(2\int_0^t |f|^2\, ds + E(z^N(0))\right). \tag{4.27}$$

The estimate (4.23) thus follows from the fact that for $z^N \in Z^N$

$$\|(A_0^N - A^N)z^N\| \leqq (1 + \mathrm{Var}\, \mu)\sqrt{2 \max(1, r)}E(z^N).$$

Hence, the lemma follows from (4.18) and (4.22).

COROLLARY 4.8. *For every* $(\phi(0), \phi) \in \mathrm{dom}\,(A_0)$ *and* f *locally square integrable,*

$$\|\kappa^N z^N(t) - z(t)\|_{\mathrm{dom}\,(A_0)} \to 0 \quad \text{as } N \to \infty,$$

where $\kappa^N : Z^N \to \mathrm{dom}\,(A_0) \subset Z$ *is defined by*

$$\kappa^N z^N = \left(a_0, \sum_{k=0}^{N} a_k L_k^N(\cdot)\right) \quad \text{with } z^N = j^N a,$$

and

$$L_k^N(\theta) = \begin{cases} -\dfrac{N}{r}(\theta - \tau_{i-1}^N) & \text{on } [\tau_i^N, \tau_{i-1}^N], \\ \dfrac{N}{r}(\theta - \tau_{i+1}^N) & \text{on } [\tau_{i+1}^N, \tau_i^N], \\ 0 & \text{elsewhere.} \end{cases}$$

Proof. The corollary follows from (4.27) and the arguments in [12, § 6].

THEOREM 4.9. *Assume* Q *and* V *are defined by* (4.21). *Let us adopt the same notation as in Theorem* 3.5. *If* (A, B) *and* (A, V) *are stabilizable,* (A, C) *and* (A, Q) *are detectable, and moreover for some* $\delta < 0$

$$\sup\{\mathrm{Re}\, \lambda : \lambda \in \sigma(A^N - \hat{G}_\delta^N C^N)\} < 0$$

for N *sufficiently large, then* $\hat{H}_\delta^N$ *defined by* (3.2) *generates a stable semigroup.*

Proof. It follows from Lemma 4.7 that (4.19) is valid when G^N is replaced by $\hat{G}_\delta^N$. Hence, the theorem can be proved combining the arguments in the proof of Theorem 4.1 with Theorem 4.2 and the formula (4.19).

Recently, we developed a higher order approximation based on linear spline elements in [11]. It possesses the exact same properties (e.g., Theorem 4.2 and Lemmas 4.4 and 4.7) as the averaging approximation and thus it satisfies Theorem 4.9.

REFERENCES

[1] M. J. BALAS, *Exponentially stabilizing finite-dimensional controller for linear distributed parameter systems: Galerkin approximation of infinite dimensional controllers*, J. Math. Anal. Appl., 117 (1986), pp. 354–368.

[2] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods on averaging approximations*, SIAM J. Control Optim., 16 (1978), pp. 169–208.

[3] H. T. BANKS AND K. KUNISCH, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684–698.

[4] R. F. CURTAIN, *Finite dimensional compensators for parabolic distributed systems with unbounded control and observation*, SIAM J. Control Optim., 22 (1984), pp. 255–276.

[5] R. F. CURTAIN AND D. SALAMON, *Finite dimensional compensators for infinite dimensional systems with unbounded input operators*, SIAM J. Control Optim., 24 (1986), pp. 797–816.

[6] J. S. GIBSON, *Linear-quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95–139.

[7] J. S. GIBSON AND A. ADAMIAN, *Approximation theory for linear quadratic Gaussian optimal control of flexible structures*, SIAM J. Control Optim., 29 (1991), to appear.

[8] T. KATO, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 1980.

[9] K. ITO, *Regulator problem for hereditary differential systems with control delays*, ICASE Report 82-3, NASA-Langley Research Center, Hampton, VA, 1982.

[10] ———, *Strong convergence and convergence rates of approximating solutions for algebraic Riccati equations in Hilbert spaces*, in Proc. Conference on Control and Identification of Distributed Parameter Systems, Vorau, July 6–12, 1986, Lecture Notes in Control and Information Sci. 102, Springer-Verlag, New York, Berlin, 1987, 153–166.

[11] K. ITO AND F. KAPPEL, *A uniformly differentiable approximation scheme for delay systems using splines*, Applied Math. Optim., to appear.

[12] K. ITO AND R. G. TEGLAS, *Legendre-tau approximation for functional differential equations*, SIAM J. Control Optim., 24 (1986), pp. 737–759.

[13] I. LASIECKA AND A. MANITIUS, *Differentiability and convergence rates of approximating semigroups for retarded functional differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 883–926.

[14] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[15] A. J. PRITCHARD AND D. SALAMON, *The linear quadratic control problem for infinite dimensional systems, part* I: *A semigroup theoretic approach for systems with unbounded input and output operators, part* II: *retarded systems with delays in control and observation*, University of Wisconsin, Madison, Tech. Sci. Report 2624, Mathematics Research Center, WI, 1984.

[16] Y. SAKAWA, *Feedback control of second order evolution equations with damping*, SIAM J. Control Optim., 22 (1984), pp. 343–361.

[17] D. SALAMON, *Control and Observation of Neutral Systems*, Research Notes in Mathematics 91, Pitman, London, 1984.

[18] ———, *Structure and stability of finite dimensional approximations for functional differential equations*, SIAM J. Control Optim., 23 (1985), pp. 928–951.

[19] J. M. SCHUMACHER, *A direct approach to compensator design for distributed parameter systems*, SIAM J. Control Optim., 21 (1983), pp. 823–836.

[20] R. E. SHOWALTER, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.

[21] G. STRANG AND G. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[22] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.

[23] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.

[24] R. B. VINTER, *Filter stability for stochastic evolution equations*, SIAM J. Control Optim., 15 (1977), pp. 465–485.

[25] D. S. BERNSTEIN AND D. C. HYLAND, *The optimal projection equations for finite-dimensional fixed-order dynamic compensation of infinite-dimensional systems*, SIAM J. Control Optim., 24 (1986), pp. 122–151.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1270-1293, November 1990

002

THE DISTURBANCE DECOUPLING PROBLEM FOR IMPLICIT LINEAR DISCRETE-TIME SYSTEMS*

A. BANASZUK†, M. KOCIĘCKI†, AND K. M. PRZYŁUSKI‡

Abstract. The disturbance decoupling problem for implicit linear discrete-time systems is studied in detail. Necessary and sufficient conditions for the problem to be solved are given. The results are obtained with the aid of the concepts of almost invariant, sliding and coasting subspaces for implicit systems.

Key words. implicit linear systems, disturbance decoupling problem, almost invariant subspaces, discrete-time systems, linear control theory

AMS(MOS) subject classifications. 93C05, 93B27, 93C30, 93C45

Introduction. Consider the implicit linear discrete-time system

$$(*)\qquad \begin{aligned} Ex_{k+1} &= Fx_k + Gu_k + Dz_k, \qquad k \geqq 0, \\ y_k &= Hx_k, \qquad k \geqq 0. \end{aligned}$$

The term z_k in the first of the equations above represents a disturbance that affects the system. We do not assume that the map E is invertible; moreover, *no* assumption on the regularity of the corresponding pencil $sE - F$ will be made. The systems of the form $(*)$ arise frequently in various applications. We mention here only [1], [11]-[13], [15], [18]-[20], [22], [31], and [32].

Our task in the present paper is to find (if possible) feedback K such that the output sequence (y_k) of the closed-loop system is not affected by the disturbance sequence (z_k). So we say that the *disturbance decoupling problem* (DDP) *is solvable for the quintuple* (E, F, G, D, H) if and only if we can find K with the property that for every disturbance sequence (z_k) there exists a sequence (x_k) such that $Ex_{k+1} = (F+GK)x_k + Dz_k$, $k \geqq 0$, and the corresponding output sequence (y_k) is identically zero. If in addition the difference equation $Ex_{k+1} = (F+GK)x_k$ has at most one solution for any initial condition, we say that the *disturbance decoupling problem with uniqueness* (DDPU) *is solvable for the quintuple* (E, F, G, D, H).

The methods used in the paper are based on our results presented in [3], [5], and [9] and also on a recent paper of Fletcher and Aasaraai [17]. It happens that the main tool for studying the disturbance decoupling problems is the concept of almost invariant subspaces introduced (for implicit systems) by Banaszuk, Kocięcki, and Przyłuski [5]. The ideas of [5] are essentially developed in the present paper. In particular, the new notion of strongly almost invariant subspace is introduced and studied in detail. We also generalize some of Willems results from [33] concerning the decomposition of an almost invariant subspace into the sum of sliding and coasting subspaces. This machinery allows us to obtain necessary and sufficient conditions for solving both disturbance decoupling problems. The obtained conditions have a simple geometric

* Received by the editors August 22, 1988; accepted for publication (in revised form) November 14, 1989. This work was performed under the auspices of the RP.I.02: "Teoria Sterowania i Optymalizacji Ciagłych układów dynamicznych i procesów dyskretnych."

† Institute of Control and Industrial Electronics, Warsaw University of Technology, Koszykowa 75, 00-662 Warszawa, Poland.

‡ Institute of Mathematics, Polish Academy of Sciences, Śniadeckich 8, P.O. Box 137, 00-950 Warszawa, Poland.

form, and are easy to check. The solutions of the problems studied in the paper are (in principle) constructive. To find a feedback map that solves the problems, we must compute some subspaces, which can be done using some recursive formulas.

It is worth noting that our formulation of DDP allows the possibility of choosing appropriately the initial condition x_0. More precisely, the initial condition x_0 is selected according to the disturbance sequence (z_k), which acts on the system. Thereby our DDP is a closed-loop version of that considered by Banaszuk, Kocięcki, and Przyłuski in [5]. Let us recall that in the case where $E = I$, the DDP studied in [5] is the same as that introduced by Willems in [34]. (Willems calls this problem "the disturbance decoupling problem with anticipation.") It seems however that the closed-loop version of this DDP which has been considered in [34] is not studied in the existing literature. In the present paper we show that when $E = I$ this DDP is solvable if and only if the classical DDP (the precise formulation of which is to be found in [36]) is solvable. In other words, when $E = I$, the DDP defined at the beginning of this section and the DDP of [36] coincide. Hence our DDP is a natural generalization of the DDP of [36] for the case of implicit systems.

We end this section by noting that some (other than our) disturbance decoupling problems have recently been formulated in [10], [25], [27], and [17]. Reference [10] states the problem (for the case when the pencil $[sE - F]$ is regular) in terms of the system transfer function and gives some sufficient conditions for the problem to be solved. References [25] and [27] consider the DDPU for a class of feedback maps that is larger than that taken into account in the present paper. In this context we should emphasize that the formulation of the DDPU given recently in [17] is essentially the same as ours. Reference [17] contains some necessary and sufficient conditions for solving the DDPU. Unfortunately, they are difficult to check since they are far from being explicit (cf. [17, § 5]). Despite this fact, [17] was very inspiring for us; it contains a result that allowed us to obtain an explicit and constructive solution of both DDP and DDPU.

1. Basic definitions and preliminary results. We begin by introducing some notation and terminology to be used throughout the paper. We shall write $\mathbb{Z}_0^+$ to denote the set of nonnegative integers. Then $\mathbb{Z}^- := \mathbb{Z} \setminus \mathbb{Z}_0^+$ and $\mathbb{Z}_0^- := \mathbb{Z}^- \cup \{0\}$. All linear spaces and linear maps considered in the paper are defined over $\mathbb{R}$. For any linear space $\mathcal{S}$, we denote by $\mathfrak{s}_0^+(\mathcal{S})$, $\mathfrak{s}_0^-(\mathcal{S})$, $\mathfrak{s}^-(\mathcal{S})$ the space of all $\mathcal{S}$-valued sequences defined on $\mathbb{Z}_0^+$, $\mathbb{Z}_0^-$, $\mathbb{Z}^-$, respectively. Let $\mathfrak{p}$ be an arbitrary space of sequences with values in a given linear space. Then we shall use the symbol $f\mathfrak{p}$ to denote the subspace of all sequences from $\mathfrak{p}$ with finite support. For instance, $f\mathfrak{s}_0^-(\mathcal{S})$ is the subspace of the space $\mathfrak{s}_0^-(\mathcal{S})$ consisting of all sequences with support bounded on the left. Let $\mathcal{S}$, $\mathcal{T}$ be linear spaces. Then $\mathcal{S} \approx \mathcal{T}$ means that $\mathcal{S}$ and $\mathcal{T}$ are isomorphic (of course, as linear spaces). The symbol Lat $(\mathcal{S})$ will stand for the lattice of all subspaces of $\mathcal{S}$ (cf. [14]). By $\mathcal{L}(\mathcal{S}, \mathcal{T})$ we shall mean the linear space of all linear maps $\mathcal{S} \to \mathcal{T}$. When $E \in \mathcal{L}(\mathcal{S}, \mathcal{T})$ and $\mathcal{S}_1 \subset \mathcal{S}$ we shall use the symbol $E|\mathcal{S}_1$ to denote the restriction of the map E to the space $\mathcal{S}_1$. Of course, $E|\mathcal{S}_1 \in \mathcal{L}(\mathcal{S}_1, \mathcal{T})$. If $A \in \mathcal{L}(\mathcal{S}, \mathcal{T})$ and $B \in \mathcal{L}(\mathcal{R}, \mathcal{T})$, then $A \times B \in \mathcal{L}(\mathcal{S} \times \mathcal{R}, \mathcal{T})$ is defined by $(A \times B)(s, r) := As + Br$, $(s, r) \in \mathcal{S} \times \mathcal{R}$. In the formulation of thesis of some results of the paper we shall write for abbreviation $\mathcal{Q} \subset \mathcal{R} \oplus \mathcal{S} \oplus \mathcal{T}$ (or similar statements). The reader should understand that then it is necessary to prove not only $\mathcal{Q} \subset \mathcal{R} + \mathcal{S} + \mathcal{T}$ but also that the spaces $\mathcal{R}$, $\mathcal{S}$, and $\mathcal{T}$ are independent.

Let $\mathcal{X}$, $\mathcal{Z}$, and $\mathcal{U}$ be fixed linear finite-dimensional spaces over $\mathbb{R}$ and $(E, F, G) \in \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{U}, \mathcal{Z})$ be a triple of linear maps. We have the following definitions.

DEFINITION 1.1. The set of all $((x_k), (u_k)) \in \mathcal{S}_0^+(\mathcal{X}) \times \mathcal{S}_0^+(\mathcal{U})$ such that the equation $Ex_{k+1} = Fx_k + Gu_k$ is satisfied on $\mathbb{Z}_0^+$ is called the *implicit linear discrete-time system defined by* (E, F, G) *on* $\mathbb{Z}_0^+$ and is denoted by $\tilde{\mathfrak{S}}(E, F, G)$.

DEFINITION 1.2. The set of all $((x_k), (u_k)) \in \mathcal{S}_0^-(\mathcal{X}) \times \mathcal{S}^-(\mathcal{U})$ such that the equation $Ex_{k+1} = Fx_k + Gu_k$ is satisfied on $\mathbb{Z}^-$ is called the *implicit linear discrete-time system defined by* (E, F, G) *on* $\mathbb{Z}^-$ and is denoted by $\tilde{\mathfrak{S}}(E, F, G)$.

Of course, the above-defined systems are linear subspaces of the corresponding spaces of sequences.

The following simple result relates systems on $\mathbb{Z}_0^+$ with those on $\mathbb{Z}^-$.

PROPOSITION 1.1. *The systems* $\tilde{\mathfrak{S}}(E, F, G)$ *and* $\tilde{\mathfrak{S}}(F, E, G)$ *are isomorphic. More precisely,* $((x_k), (u_k)) \in \tilde{\mathfrak{S}}(E, F, G)$ *if and only if* $((x_{-k}), (u_{-k-1})) \in \tilde{\mathfrak{S}}(F, E, G)$.

Let $\mathcal{W}$ be a given subspace of $\mathcal{X}$. Then we have the following definition.

DEFINITION 1.3. The *trace of the system* $\tilde{\mathfrak{S}}(E, F, G)$ *on* $\mathcal{W}$, to be denoted by $\tilde{\mathfrak{S}}(E, F, G) | \mathcal{W}$, is the set of all $((x_k), (u_k)) \in \tilde{\mathfrak{S}}(E, F, G)$ such that $(x_k) \in \mathcal{S}_0^+(\mathcal{W})$. (The *trace of the system* $\tilde{\mathfrak{S}}(E, F, G)$ *on* $\mathcal{W}$ is defined similarly and will be denoted by $\tilde{\mathfrak{S}}(E, F, G) | \mathcal{W}$.)

Of course, $\tilde{\mathfrak{S}}(E, F, G) | \mathcal{W} = \tilde{\mathfrak{S}}(E | \mathcal{W}, F | \mathcal{W}, G)$, i.e., the trace of a system on any subspace is a system.

In the sequel we shall need the following subspaces of $\mathcal{X}$.

DEFINITION 1.4. The *space of admissible initial conditions of the system* $\tilde{\mathfrak{S}}(E, F, G)$, to be denoted by $\vec{\mathcal{V}}(E, F, G)$, is the set of all $x \in \mathcal{X}$ for which there exists $((x_k), (u_k)) \in \tilde{\mathfrak{S}}(E, F, G)$ such that $x_0 = x$.

DEFINITION 1.5. The *space of admissible final conditions of the system* $\tilde{\mathfrak{S}}(E, F, G)$, to be denoted by $\overleftarrow{\mathcal{V}}(E, F, G)$, is the set of all $x \in \mathcal{X}$ for which there exists $((x_k), (u_k)) \in \tilde{\mathfrak{S}}(E, F, G)$ such that $x_0 = x$.

DEFINITION 1.6. The *reachable space of the system* $\tilde{\mathfrak{S}}(E, F, G)$, to be denoted by $\vec{\mathcal{R}}(E, F, G)$, is the set of all $x \in \mathcal{X}$ for which there exists $((x_k), (u_k)) \in \not{\mathfrak{S}}(E, F, G)$ such that $x_0 = x$.

DEFINITION 1.7. The *controllable space of the system* $\tilde{\mathfrak{S}}(E, F, G)$, to be denoted by $\vec{\mathcal{C}}(E, F, G)$, is defined as $\vec{\mathcal{V}}(E, F, G) \cap \vec{\mathcal{R}}(E, F, G)$.

DEFINITION 1.8. Let $i \in \mathbb{Z}_0^+$; then the *ith controllable space of the system* $\tilde{\mathfrak{S}}(E, F, G)$ is the set of all $x \in \mathcal{X}$ for which there exists $((x_k), (u_k)) \in \tilde{\mathfrak{S}}(E, F, G)$ such that $x_0 = 0$ and $x_i = x$. The space will be denoted by $\vec{\mathcal{C}}_i(E, F, G)$.

Remark 1.1. The above-introduced spaces have recently been studied in [3]-[5], and [7]. Let us observe that the space $\vec{\mathcal{C}}(E, F, G)$ can also be expressed (cf. [3, § 5]) as the set of all vectors $x \in \mathcal{X}$ such that we can find $((x_k), (u_k)) \in \tilde{\mathfrak{S}}(E, F, G)$ with the property that $x_0 = 0$ and $x_i = x$, for some $i \in \mathbb{Z}_0^+$. In other words, $\vec{\mathcal{C}}(E, F, G) = \bigcup \vec{\mathcal{C}}_i(E, F, G)$, where the summation is taken over all $i \in \mathbb{Z}_0^+$. For some supplementary remarks concerning the spaces $\vec{\mathcal{V}}(E, F, G)$, $\overleftarrow{\mathcal{V}}(E, F, G)$, $\vec{\mathcal{R}}(E, F, G)$, and $\vec{\mathcal{C}}(E, F, G)$ see the Appendix.

Let $\mathcal{W}$ be any subspace of $\mathcal{X}$. Then replacing in the definitions of $\vec{\mathcal{V}}(E, F, G)$, $\overleftarrow{\mathcal{V}}(E, F, G)$, $\vec{\mathcal{R}}(E, F, G)$, $\vec{\mathcal{C}}(E, F, G)$, and $\vec{\mathcal{C}}_i(E, F, G)$ the systems $\tilde{\mathfrak{S}}(E, F, G)$ and $\tilde{\mathfrak{S}}(E, F, G)$ by their traces on $\mathcal{W}$ we obtain the subspaces of $\mathcal{X}$, which will be denoted by $\vec{\mathcal{V}}(E, F, G) | \mathcal{W}$, $\overleftarrow{\mathcal{V}}(E, F, G) | \mathcal{W}$, $\vec{\mathcal{R}}(E, F, G) | \mathcal{W}$, $\vec{\mathcal{C}}(E, F, G) | \mathcal{W}$, and $\vec{\mathcal{C}}_i(E, F, G) | \mathcal{W}$, respectively. The relationship between the above-introduced spaces and related spaces considered in the existing literature is presented in the Appendix.

In the sequel we will need the following definitions.

DEFINITION 1.9. Let $\mathcal{W} \subset \mathcal{X}$. We say that $\mathcal{W}$ is $\tilde{\mathfrak{S}}(E, F, G)$-*invariant* if and only if $\mathcal{W} = \vec{\mathcal{V}}(E, F, G) | \mathcal{W}$.

DEFINITION 1.10. Let $\mathcal{W} \subset \mathcal{X}$. We say that $\mathcal{W}$ is $\tilde{\mathfrak{S}}(E, F, G)$-*invariant* if and only if $\mathcal{W} = \overleftarrow{\mathcal{V}}(E, F, G) | \mathcal{W}$.

DEFINITION 1.11. Let $\mathcal{W} \subset \mathcal{X}$. We say that $\mathcal{W}$ is a *reachability subspace for the system* $\vec{\mathfrak{S}}(E, F, G)$ if and only if $\mathcal{W} = \vec{\mathcal{R}}(E, F, G)|\mathcal{W}$.

DEFINITION 1.12. Let $\mathcal{W} \subset \mathcal{X}$. We say that $\mathcal{W}$ is a *controllability subspace for the system* $\vec{\mathfrak{S}}(E, F, G)$ if and only if $\mathcal{W} = \vec{\mathcal{C}}(E, F, G)|\mathcal{W}$.

DEFINITION 1.13. Let $\mathcal{W} \subset \mathcal{X}$. We say that $\mathcal{W}$ is a *uniqueness subspace for the system* $\vec{\mathfrak{S}}(E, F, G)$ if and only if $\vec{\mathcal{C}}(E, F, G)|\mathcal{W} = 0$.

(We can easily check that $\mathcal{W}$ is a uniqueness subspace for the system if and only if $((x_k^i), (u_k^i)) \in \vec{\mathfrak{S}}(E, F, G)|\mathcal{W}$, for $i = 1, 2$, and $x_0^1 = x_0^2$ implies $(x_k^1) = (x_k^2)$.)

The most important properties of $\vec{\mathfrak{S}}(E, F, G)$- and $\overleftarrow{\mathfrak{S}}(E, F, G)$-invariant subspaces are collected in the following proposition (cf. [3, §§ 2, 5]).

PROPOSITION 1.2. *Let* $(E, F, G) \in \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{U}, \mathcal{Z})$ *be given and* $\mathcal{W}$ *be any subspace of* $\mathcal{X}$. *Then the following statements hold*:

(1) $\mathcal{W}$ *is* $\vec{\mathfrak{S}}(E, F, G)$*-invariant if and only if* $F\mathcal{W} \subset E\mathcal{W} + \operatorname{Im} G$ (*equivalently*, $\mathcal{W} \subset F^{-1}(E\mathcal{W} + \operatorname{Im} G)$).

(2) $\mathcal{W}$ *is* $\overleftarrow{\mathfrak{S}}(E, F, G)$*-invariant if and only if* $\mathcal{W}$ *is* $\vec{\mathfrak{S}}(F, E, G)$*-invariant.*

(3) *The sum of any number of* $\vec{\mathfrak{S}}(E, F, G)$*-invariant subspaces is also* $\vec{\mathfrak{S}}(E, F, G)$*-invariant.*

(4) $\vec{\mathcal{V}}(E, F, G)$ *is* $\vec{\mathfrak{S}}(E, F, G)$*-invariant; moreover,* $\vec{\mathcal{V}}(E, F, G) = F^{-1}(E\vec{\mathcal{V}}(E, F, G) + \operatorname{Im} G)$.

(5) $\vec{\mathcal{R}}(E, F, G)$ *is* $\overleftarrow{\mathfrak{S}}(E, F, G)$*-invariant.*

(6) $\vec{\mathcal{C}}(E, F, G)$ *is* $\vec{\mathfrak{S}}(E, F, G)$*- and* $\overleftarrow{\mathfrak{S}}(E, F, G)$*-invariant*; *moreover,* $\vec{\mathcal{C}}(E, F, G) = E^{-1}(F\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G) \cap \vec{\mathcal{V}}(E, F, G)$.

The result below is an immediate consequence of Proposition 1.2(1).

PROPOSITION 1.3. *Let* $\mathcal{W}$ *be any* $\vec{\mathfrak{S}}(E, F, G)$*-invariant subspace. Assume* $\mathcal{S} \subset \mathcal{X}$ *satisfies* $\mathcal{S} \oplus (\mathcal{W} \cap \operatorname{Ker} E) = \mathcal{W}$. *Then* $\mathcal{S}$ *is* $\vec{\mathfrak{S}}(E, F, G)$*-invariant and* $E\mathcal{S} \approx \mathcal{S}$.

In the present paper we shall also need the following properties of the spaces $\vec{\mathcal{C}}(E, F, G)$ and $\vec{\mathcal{C}}_j(E, F, G)$ (cf. [3, § 5] and [4]).

PROPOSITION 1.4. *Let* $(E, F, G) \in \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{U}, \mathcal{Z})$ *be given. Then the following statements hold*:

(1) *Let* $i, j \in \mathbb{Z}_0^+$. *Then* $\vec{\mathcal{C}}_i(E, F, G) \subset \vec{\mathcal{C}}_j(E, F, G)$ *if* $i \leqq j$.

(2) *There exists* $\gamma \in \mathbb{Z}_0^+$ *such that* $\vec{\mathcal{C}}_i(E, F, G) = \vec{\mathcal{C}}(E, F, G)$, *for* $i \geqq \gamma$.

(3) $\vec{\mathcal{C}}(E, F, G)$ *is the smallest* $\vec{\mathfrak{S}}(E, F, G)$*-invariant subspace containing* $\vec{\mathcal{C}}_1(E, F, G)$.

(4) $\vec{\mathcal{C}}(E, F, G) = \vec{\mathcal{C}}(F, E, G)$.

(5) *For any* $i \in \mathbb{Z}_0^+$, $E\vec{\mathcal{C}}_{i+1}(E, F, G) + \operatorname{Im} G = F\vec{\mathcal{C}}_i(E, F, G) + \operatorname{Im} G$. *In particular,* $E\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G = F\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G$.

The smallest nonnegative integer γ for which the statement (2) of the above proposition holds will be denoted by $\gamma(E, F, G)$. (In [3] the integer $\gamma(E, F, G)$ is called "the *controllability index of the system* $\vec{\mathfrak{S}}(E, F, G)$.")

DEFINITION 1.14. The set of all $((x_k), (u_k), (z_k)) \in \mathcal{S}_0^+(\mathcal{X}) \times \mathcal{S}_0^+(\mathcal{U}) \times \mathcal{S}_0^+(\mathcal{Z})$ such that the equation $Ex_{k+1} = Fx_k + Gu_k + z_k$ holds on $\mathbb{Z}_0^+$ will be called the *implicit linear discrete-time system with disturbances defined by* (E, F, G) *on* $\mathbb{Z}_0^+$ and will be denoted by $\vec{\mathfrak{S}}_d(E, F, G)$.

We can prove that the system $\vec{\mathfrak{S}}_d(E, F, G)$ (meant as a subspace of $\mathcal{S}_0^+(\mathcal{X}) \times \mathcal{S}_0^+(\mathcal{U}) \times \mathcal{S}_0^+(\mathcal{Z})$) determines uniquely (in opposition to the system $\vec{\mathfrak{S}}(E, F, G)$) the triple (E, F, G).

DEFINITION 1.15. Let $\mathcal{D} \subset \mathcal{Z}$. We say (cf. [3], [5]) that the system $\vec{\mathfrak{S}}_d(E, F, G)$ *accepts all disturbance sequences from* $\mathcal{S}_0^+(\mathcal{D})$ if and only if for every $(z_k) \in \mathcal{S}_0^+(\mathcal{D})$ there exists $((x_k), (u_k)) \in \mathcal{S}_0^+(\mathcal{X}) \times \mathcal{S}_0^+(\mathcal{U})$ such that $((x_k), (u_k), (z_k)) \in \vec{\mathfrak{S}}_d(E, F, G)$.

For a system $\vec{\mathfrak{S}}_d(E, F, G)$ and a subspace $\mathcal{D} \subset \mathcal{Z}$, we shall denote by $\lambda(E, F, G; \mathcal{D})$ the greatest lower bound of the set of $\lambda \in \mathbb{Z}_0^+$ for which the following implication holds:

$$\forall (z_k)\in \mathcal{S}_0^+(\mathcal{D}):\quad [((z_k)_{k=0}^{\lambda-1}=0)\Rightarrow(\exists((x_k),(u_k))\in \mathcal{S}_0^+(\mathcal{X})\times\mathcal{S}_0^+(\mathcal{U}):$$
$$x_0=0 \text{ and } ((x_k),(u_k),(z_k))\in \tilde{\mathfrak{S}}_d(E,F,G))].$$

To study some properties of a system with disturbances we will need the following remark, which shows that such a system can be conveniently analyzed with the aid of a properly defined system without disturbances. It will allow us to use results of [3].

Remark 1.2. Let $D\in\mathcal{L}(\mathcal{D},\mathcal{X})$ be the canonical injection $\mathcal{D}\to\mathcal{X}$ and $Q\in\mathcal{L}(\mathcal{X},\mathcal{X}/\operatorname{Im} G)$ be the canonical surjection $\mathcal{X}\to\mathcal{X}/\operatorname{Im} G$. Let $(\tilde{E},\tilde{F},\tilde{D}):=(QE,QF,QD)$. Observe now that if $((x_k),(z_k))\in\tilde{\mathfrak{S}}(\tilde{E},\tilde{F},\tilde{D})$, then $(z_k)\in\mathcal{S}_0^+(\mathcal{D})$ and there exists $(u_k)\in\mathcal{S}_0^+(\mathcal{U})$ such that $((x_k),(u_k),(z_k))\in\tilde{\mathfrak{S}}_d(E,F,G)$. Conversely, let $(z_k)\in\mathcal{S}_0^+(\mathcal{D})$ and $((x_k),(u_k),(z_k))\in\tilde{\mathfrak{S}}_d(E,F,G)$. Then $((x_k),(z_k))\in\tilde{\mathfrak{S}}(\tilde{E},\tilde{F},\tilde{D})$. It follows that the system $\tilde{\mathfrak{S}}_d(E,F,G)$ accepts all disturbance sequences from $\mathcal{S}_0^+(\mathcal{D})$ if and only if the system $\tilde{\mathfrak{S}}_d(\tilde{E},\tilde{F},0)$ accepts all disturbance sequences from $\mathcal{S}_0^+(\operatorname{Im}\tilde{D})$. Hence the system $\tilde{\mathfrak{S}}_d(E,F,G)$ accepts all disturbance sequences from $\mathcal{S}_0^+(\mathcal{D})$ if and only if the system $\tilde{\mathfrak{S}}(\tilde{E},\tilde{F},\tilde{D})$ accepts all input sequences in the sense of [3, § 4]. Another consequence of the relation between the system $\tilde{\mathfrak{S}}_d(E,F,G)$ and $\tilde{\mathfrak{S}}(\tilde{E},\tilde{F},0)$ is that $\vec{\mathcal{V}}(E,F,G)=\vec{\mathcal{V}}(\tilde{E},\tilde{F},0)$. Similarly (considering the system $\tilde{\mathfrak{S}}(\tilde{E},\tilde{F},\tilde{D})$), we can observe that $\vec{\mathcal{R}}(E,F,G)=\vec{\mathcal{R}}(\tilde{E},\tilde{F},0)$.

Let us note also that, in view of Corollary 3.3 and Proposition 4.1 of [3] if a system $\tilde{\mathfrak{S}}_d(E,F,G)$ accepts all disturbance sequences from $\mathcal{S}_0^+(\mathcal{D})$, then $\lambda(E,F,G;\mathcal{D})$ coincides with the anticipation index of the system $\tilde{\mathfrak{S}}(\tilde{E},\tilde{F},\tilde{D})$, the quantity being studied in [3, § 3]. We refer the interested reader to [3, § 3] for further remarks concerning the anticipation phenomenon.

Now we can summarize basic properties of systems with disturbances.

PROPOSITION 1.5. *Let $\mathcal{D}\subset\mathcal{X}$ be arbitrary. Then the following statements are equivalent*:

(a) $\tilde{\mathfrak{S}}_d(E,F,G)$ *accepts all disturbance sequences from* $\mathcal{S}_0^+(\mathcal{D})$.

(b) $\mathcal{D}\subset E\vec{\mathcal{V}}(E,F,G)+F\vec{\mathcal{R}}(E,F,G)+\operatorname{Im} G$.

(c) $\lambda(E,F,G;\mathcal{D})$ *is infinite.*

Proof. Let $(\tilde{E},\tilde{F},\tilde{D})$ be defined as in Remark 1.2.

(a)$\Leftrightarrow$(b) Since the inclusion $\mathcal{D}\subset E\vec{\mathcal{V}}(E,F,G)+F\vec{\mathcal{R}}(E,F,G)+\operatorname{Im} G$ is obviously equivalent to the inclusion $\operatorname{Im}\tilde{D}\subset\tilde{E}\vec{\mathcal{V}}(\tilde{E},\tilde{F},0)+\tilde{F}\vec{\mathcal{R}}(\tilde{E},\tilde{F},0)$, the equivalence follows from Theorem 4.1 and Proposition 2.7 and 2.8 of [3].

(a)$\Rightarrow$(c) In view of Remark 1.2, $\lambda(E,F,G;\mathcal{D})$ coincides with the anticipation index of the system $\tilde{\mathfrak{S}}(\tilde{E},\tilde{F},\tilde{D})$. Since $\dim\mathcal{X}<\infty$, it follows from Corollary 3.3 of [3] that $\lambda(E,F,G;\mathcal{D})$ is finite.

(c)$\Rightarrow$(a) Consider an arbitrary $(z_k)\in\mathcal{S}_0^+(\mathcal{D})$. Let λ be an integer not less than $\lambda(E,F,G;\mathcal{D})$, and let $(\bar{z}_k)\in\mathcal{S}_0^+(\mathcal{D})$ be defined as $\bar{z}_k:=0$ for $k=0,\cdots,\lambda-1$, and $\bar{z}_k:=z_{k-\lambda}$ for $k=\lambda,\lambda+1,\cdots$. Then there exists $((\bar{x}_k),(\bar{u}_k))\in\mathcal{S}_0^+(\mathcal{X})\times\mathcal{S}_0^+(\mathcal{U})$ such that $((\bar{x}_k),(\bar{u}_k),(\bar{z}_k))\in\tilde{\mathfrak{S}}_d(E,F,G)$. Put $(x_k,u_k):=(\bar{x}_{k+\lambda},\bar{u}_{k+\lambda})$ for $k\in\mathbb{Z}_0^+$. Now it is sufficient to note that $((x_k),(u_k),(z_k))\in\tilde{\mathfrak{S}}_d(E,F,G)$. □

We now record the following lemma.

LEMMA 1.1. *Let $\mathcal{D}\subset E\vec{\mathcal{V}}(E,F,G)+\operatorname{Im} G$. Then $\lambda(E,F,G;\mathcal{D})=0$ and hence $\tilde{\mathfrak{S}}_d(E,F,G)$ accepts all disturbance sequences from $\mathcal{S}_0^+(\mathcal{D})$.*

Proof. In view of Remark 1.2 and the proof of Proposition 1.5, the result follows from Theorem 4.1 and Proposition 2.7 of [3]. □

The concept of almost $\tilde{\mathfrak{S}}(E,F,G)$-invariant subspace has been recently introduced in [5]. Let us recall the definition from [5].

DEFINITION 1.16. Let $\mathcal{W}\subset\mathcal{X}$. We say that $\mathcal{W}$ is *almost $\tilde{\mathfrak{S}}(E,F,G)$-invariant* if and only if for any $(z_k)\in\mathcal{S}_0^+(E\mathcal{W}+F\mathcal{W}+\operatorname{Im} G)$ there exists $((x_k),(u_k))\in\mathcal{S}_0^+(\mathcal{X})\times\mathcal{S}_0^+(\mathcal{U})$ such that $((x_k),(u_k),(z_k))\in\tilde{\mathfrak{S}}_d(E,F,G)|\mathcal{W}$.

Remark 1.3. We shall need in the sequel some results of [5]. The results concern a system with disturbances described by the difference equation

$$(*) \qquad Ex_{k+1} = Fx_k + Gu_k + z_k,$$

on $\mathbb{Z}$. So the theory of [5] is not directly applicable in the setting of the present paper. In order to use the results of [5] to the system $\tilde{\mathfrak{S}}_d(E, F, G)$ it is sufficient to note that there exists a natural correspondence between solutions of the equation (*) with support bounded on the left and some solutions of this equation on $\mathbb{Z}_0^+$. More precisely, let $((x_k)_{k=-\infty}^{\infty}, (u_k)_{k=-\infty}^{\infty}, (z_k)_{k=-\infty}^{\infty})$ be a solution of (*) satisfying $(x_k)_{k=-\infty}^{\alpha} = 0$, $(u_k)_{k=-\infty}^{\alpha-1} = 0$, $(z_k)_{k=-\infty}^{\alpha-1} = 0$ for some $\alpha \in \mathbb{Z}$. Let $((\tilde{x}_k)_{k=0}^{\infty}, (\tilde{u}_k)_{k=0}^{\infty}, (\tilde{z}_k)_{k=0}^{\infty}) := ((x_{k+\alpha})_{k=0}^{\infty}, (u_{k+\alpha})_{k=0}^{\infty}, (z_{k+\alpha})_{k=0}^{\infty})$. Then $((\tilde{x}_k)_{k=0}^{\infty}, (\tilde{u}_k)_{k=0}^{\infty}, (\tilde{z}_k)_{k=0}^{\infty})$ is a solution of (*) on $\mathbb{Z}_0^+$ satisfying $\tilde{x}_0 = 0$. Conversely, let $((\tilde{x}_k)_{k=0}^{\infty}, (\tilde{u}_k)_{k=0}^{\infty}, (\tilde{z}_k)_{k=0}^{\infty})$ be a solution of (*) on $\mathbb{Z}_0^+$ with $\tilde{x}_0 = 0$. Put $((x_k)_{k=-\infty}^{\alpha-1}, (u_k)_{k=-\infty}^{\alpha-1}, (z_k)_{k=-\infty}^{\alpha-1}) := 0$ and $((x_k)_{k=\alpha}^{\infty}, (u_k)_{k=\alpha}^{\infty}, (z_k)_{k=\alpha}^{\infty}) := ((\tilde{x}_k)_{k=0}^{\infty}, (\tilde{u}_k)_{k=0}^{\infty}, (\tilde{z}_k)_{k=0}^{\infty})$. Then it is immediate that $((x_k)_{k=-\infty}^{\infty}, (u_k)_{k=-\infty}^{\infty}, (z_k)_{k=-\infty}^{\infty})$ is a solution of (*) with support bounded on the left. The above considerations allow us to apply directly some results of [5] to study properties of the system $\tilde{\mathfrak{S}}_d(E, F, G)$. In particular, it allows us to identify the notion of almost invariant subspace introduced above with that considered in [5].

The following proposition (cf. [5, Thm. 3.2]) gives a geometric characterization of almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant subspaces.

PROPOSITION 1.6. *Let $\mathcal{W}$ be a subspace of $\mathcal{X}$. Then the following statements are equivalent*:

(a) *$\mathcal{W}$ is almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant.*

(b) $\mathcal{W} = \vec{V}(E, F, G)|\mathcal{W} + \vec{\mathcal{R}}(E, F, G)|\mathcal{W}$.

(c) $E\mathcal{W} + F\mathcal{W} + \operatorname{Im} G = E\vec{V}(E, F, G)|\mathcal{W} + F\vec{\mathcal{R}}(E, F, G)|\mathcal{W} + \operatorname{Im} G$.

Using the above proposition we can prove (cf. [5]) the following properties of almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant subspaces.

COROLLARY 1.1. (1) *Every $\tilde{\mathfrak{S}}(E, F, G)$-invariant (or $\tilde{\mathfrak{S}}(E, F, G)$-invariant) subspace is almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant. In particular, the spaces $\vec{V}(E, F, G)$, $\vec{V}(E, F, G)$, $\vec{\mathcal{R}}(E, F, G)$, and $\vec{\mathcal{C}}(E, F, G)$ are almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant.*

(2) *The sum of any number of almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant subspaces is also almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant.*

(3) *There exists a greatest almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant subspace. The subspace coincides with $\vec{V}(E, F, G) + \vec{\mathcal{R}}(E, F, G)$.*

Let us recall (cf. [3], [11], [16], [35]) the following definitions.

DEFINITION 1.17. Let $\tilde{\mathfrak{S}}_d(E, F, G)$ be a given system and let $(x_k) \in \mathcal{S}_0^+(\mathcal{X})$ be such that $x_0 = 0$. Then we will say that the system $\tilde{\mathfrak{S}}_d(E, F, G)$ *possesses the uniqueness property* if and only if $((x_k), 0, 0) \in \tilde{\mathfrak{S}}_d(E, F, G)$ implies $(x_k) = 0$.

DEFINITION 1.18. We say (cf. [3]) that a system $\tilde{\mathfrak{S}}_d(E, F, G)$ is *regular* if and only if the system possesses the uniqueness property and the system $\tilde{\mathfrak{S}}_d(E, F, 0)$ accepts all disturbance sequences.

(Of course, the properties defined above do not depend on the map G.)

We record here the following result concerning regular systems (cf. [3, Remark 4.2, Props. 4.13, 4.14]).

PROPOSITION 1.7. *Let $(E, F, G) \in \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{U}, \mathcal{Z})$ be a given triple of linear maps. Then the following statements are equivalent*:

(a) *$\tilde{\mathfrak{S}}_d(E, F, G)$ is regular.*

(b) *The pencil $[sE - F]$ is regular.*

(c) *$\dim \mathcal{Z} = \dim \mathcal{X}$ and $\tilde{\mathfrak{S}}_d(E, F, G)$ possesses the uniqueness property.*

(d) *$\dim \mathcal{Z} = \dim \mathcal{X}$ and $\tilde{\mathfrak{S}}_d(E, F, 0)$ accepts all disturbance sequences.*

Another consequence of Proposition 4.13 of [3] is the following result.

PROPOSITION 1.8. $\dim \mathscr{Z} \leqq \dim \mathscr{X}$, *if* $\tilde{\mathfrak{S}}_d(E, F, 0)$ *accepts all disturbance sequences.*

2. On certain decompositions. In this section we investigate some decompositions of the implicit system. We will also study relationships between various properties of such a system and analogous properties of its subsystems.

Let $\mathscr{X} = \mathscr{X}_1 \oplus \cdots \oplus \mathscr{X}_p$, $\mathscr{Z} = \mathscr{Z}_1 \oplus \cdots \oplus \mathscr{Z}_p$, and $\mathscr{U} = \mathscr{U}_1 \oplus \cdots \oplus \mathscr{U}_q$, for some positive integers p and q and assume that $(E, F, G) \in \mathscr{L}(\mathscr{X}, \mathscr{Z}) \times \mathscr{L}(\mathscr{X}, \mathscr{Z}) \times \mathscr{L}(\mathscr{U}, \mathscr{Z})$ is fixed. We introduce the following definitions.

DEFINITION 2.1. The $(2p+q)$-tuple of subspaces $(\mathscr{X}_1, \cdots, \mathscr{X}_p; \mathscr{Z}_1, \cdots, \mathscr{Z}_p; \mathscr{U}_1, \cdots, \mathscr{U}_q)$ is said to be a *decomposition of* $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$. A decomposition of $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$ is said to be *good for the triple* (E, F, G) if and only if $\mathscr{Z}_1 = E\mathscr{X}_1 + F\mathscr{X}_1 + G\mathscr{U}_1$. In the particular case when $\mathscr{X}_1 = \vec{\mathscr{C}}(E, F, G)$ and $q = 1$ we say that the corresponding good decomposition is a *Kalman decomposition* (cf. [9]).

(Note that in the case of a Kalman decomposition we have $\mathscr{Z}_1 = E\vec{\mathscr{C}}(E, F, G) + \operatorname{Im} G = F\vec{\mathscr{C}}(E, F, G) + \operatorname{Im} G$; cf. Proposition 1.4(5).)

DEFINITION 2.2. Let $J_i \in \mathscr{L}(\mathscr{X}_i, \mathscr{X})$ be the canonical injection $\mathscr{X}_i$ into $\mathscr{X}$, let $M_m \in \mathscr{L}(\mathscr{U}_m, \mathscr{U})$ be the canonical injection of $\mathscr{U}_m$ into $\mathscr{U}$, and let $P_i \in \mathscr{L}(\mathscr{Z}, \mathscr{Z}_i)$ denote the canonical surjection of $\mathscr{Z}$ onto $\mathscr{Z}_i$. Define, for $i, j = 1, \cdots, p$, $m = 1, \cdots, q$, $E_{ij} \in \mathscr{L}(\mathscr{X}_j, \mathscr{Z}_i)$, $F_{ij} \in \mathscr{L}(\mathscr{X}_j, \mathscr{Z}_i)$, $G_{im} \in \mathscr{L}(\mathscr{U}_m, \mathscr{Z}_i)$ by $E_{ij} := P_i E J_j$, $F_{ij} := P_i F J_j$, and $G_{im} := P_i G M_m$. The (ordered) $(2p^2+pq)$-tuple of linear maps $(E_{11}, E_{12}, \cdots, E_{pp}; F_{11}, F_{12}, \cdots, F_{pp}; G_{11}, G_{12}, \cdots, G_{pq})$ is called the $(2p^2+pq)$*-tuple corresponding to the triple* (E, F, G) *for the decomposition* $(\mathscr{X}_1, \cdots, \mathscr{X}_p; \mathscr{Z}_1, \cdots, \mathscr{Z}_p; \mathscr{U}_1, \cdots, \mathscr{U}_q)$ *of* $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$.

Let us observe that if the decomposition $(\mathscr{X}_1, \cdots, \mathscr{X}_p; \mathscr{Z}_1, \cdots, \mathscr{Z}_p; \mathscr{U}_1, \cdots, \mathscr{U}_q)$ is good for (E, F, G), then $E_{i1} = F_{i1} = 0$ and $G_{i1} = 0$, for all $i \neq 1$.

The following propositions are useful.

PROPOSITION 2.1. *Let* $(\mathscr{X}_1, \cdots, \mathscr{X}_p; \mathscr{Z}_1, \cdots, \mathscr{Z}_p; \mathscr{U})$ *be a good decomposition of* $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$ *for* (E, F, G), *and let* $(E_{11}, E_{12}, \cdots, E_{pp}; F_{11}, F_{12}, \cdots, F_{pp}; G_1, \cdots, G_p)$ *be the* $(2p^2+p)$*-tuple corresponding to* (E, F, G) *for this decomposition. Then the system* $\tilde{\mathfrak{S}}_d(E_{11}, F_{11}, G_1)$ *accepts all disturbance sequences from* $\mathscr{S}_0^+(\mathscr{Z}_1)$ *if and only if* $\mathscr{X}_1$ *is an almost* $\tilde{\mathfrak{S}}(E, F, G)$*-invariant subspace.*

Proof. The proof follows immediately from the corresponding definitions. □

PROPOSITION 2.2. *Let* $(\mathscr{X}_1, \mathscr{X}_2; \mathscr{Z}_1, \mathscr{Z}_2; \mathscr{U})$ *be a good decomposition of* $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$ *for* (E, F, G), *and let* $(E_{11}, E_{12}, 0, E_{22}; F_{11}, F_{12}, 0, F_{22}; G_1, 0)$ *be the* 10*-tuple corresponding to* (E, F, G) *for this decomposition. Let* $\mathscr{D} \subset \mathscr{Z}$ *and* $\mathscr{D}_2 \subset \mathscr{Z}_2$ *satisfy* $\mathscr{D}_2 \subset \mathscr{D}$ *and* $\mathscr{Z}_1 \oplus \mathscr{D}_2 = \mathscr{Z}_1 + \mathscr{D}$. *Then the system* $\tilde{\mathfrak{S}}_d(E_{22}, F_{22}, 0)$ *accepts all disturbance sequences from* $\mathscr{S}_0^+(\mathscr{D}_2)$ *if the system* $\tilde{\mathfrak{S}}_d(E, F, G)$ *accepts all disturbance sequences from* $\mathscr{S}_0^+(\mathscr{D})$. *Conversely, the system* $\tilde{\mathfrak{S}}_d(E, F, G)$ *accepts all disturbance sequences from* $\mathscr{S}_0^+(\mathscr{D})$ *if the system* $\tilde{\mathfrak{S}}_d(E_{11}, F_{11}, G_1)$ *accepts all disturbance sequences from* $\mathscr{S}_0^+(\mathscr{Z}_1)$ *and the system* $\tilde{\mathfrak{S}}_d(E_{22}, F_{22}, 0)$ *accepts all disturbance sequences from* $\mathscr{S}_0^+(\mathscr{D}_2)$.

Proof. The first statement is obvious. To show the second statement consider an arbitrary $(z_k) \in \mathscr{S}_0^+(\mathscr{D})$. Then there exist $(z_k^1) \in \mathscr{S}_0^+(\mathscr{Z}_1)$ and $(z_k^2) \in \mathscr{S}_0^+(\mathscr{D}_2)$ such that $(z_k) = (z_k^1) + (z_k^2)$. Since $\tilde{\mathfrak{S}}_d(E_{22}, F_{22}, 0)$ accepts all disturbance sequences from $\mathscr{S}_0^+(\mathscr{D}_2)$ we can find $(x_k^2) \in \mathscr{S}_0^+(\mathscr{X}_2)$ such that $((x_k^2), 0, (z_k^2)) \in \tilde{\mathfrak{S}}_d(E_{22}, F_{22}, 0)$. Similarly, the system $\tilde{\mathfrak{S}}_d(E_{11}, F_{11}, G_1)$ accepts all disturbance sequences from $\mathscr{S}_0^+(\mathscr{Z}_1)$; hence there exist $((x_k^1), (u_k)) \in \mathscr{S}_0^+(\mathscr{X}_1) \times \mathscr{S}_0^+(\mathscr{U})$ such that $((x_k^1), (u_k), (-E_{12}x_{k+1}^2 + F_{12}x_k^2 + z_k^2)) \in \tilde{\mathfrak{S}}_d(E_{11}, F_{11}, G_1)$. Thus $((x_k^1 + x_k^2), (u_k), (z_k)) \in \tilde{\mathfrak{S}}_d(E, F, G)$. □

PROPOSITION 2.3. *Let* $(\mathscr{X}_1, \mathscr{X}_2; \mathscr{Z}_1, \mathscr{Z}_2; \mathscr{U})$ *be a good decomposition of* $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$

for (E, F, G), *and let* $(E_{11}, E_{12}, 0, E_{22}; F_{11}, F_{12}, 0, F_{22}; G_1, 0)$ *be the* 10*-tuple corresponding to* (E, F, G) *for this decomposition. Then the system* $\vec{\mathfrak{S}}_d(E, F, G)$ *accepts all disturbance sequences from* $\mathfrak{s}_0^+(\mathscr{X})$ *if the systems* $\vec{\mathfrak{S}}_d(E_{11}, F_{11}, G_1)$ *and* $\vec{\mathfrak{S}}_d(E_{22}, F_{22}, 0)$ *accept all disturbance sequences from* $\mathfrak{s}_0^+(\mathscr{X}_1)$ *and* $\mathfrak{s}_0^+(\mathscr{X}_2)$, *respectively.*

Proof. The proof follows from the second statement of Proposition 2.2. For $\mathscr{D} = \mathscr{X}$ and $\mathscr{D}_2 = \mathscr{X}_2$. □

In § 3 we shall need the following result.

PROPOSITION 2.4. *Let* $(\mathscr{X}_1, \mathscr{X}_2; \mathscr{Z}_1, \mathscr{Z}_2; \mathscr{U}_1, \mathscr{U}_2)$ *be a good decomposition of* $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$ *for* (E, F, G), *and let* $(E_{11}, E_{12}, 0, E_{22}; F_{11}, F_{12}, 0, F_{22}; G_{11}, G_{12}, 0, G_{22})$ *be the* 12*-tuple corresponding to* (E, F, G) *for this decomposition. Assume in addition that* $\mathscr{X}_1 = \vec{\mathscr{C}}(E_{11}, F_{11}, G_{11})$. *Then* $\vec{\mathscr{C}}(E, F, G) = \mathscr{X}_1 \oplus \vec{\mathscr{C}}(E_{22}, F_{22}, G_{22})$ *and* $E\vec{\mathscr{C}}(E, F, G) + \operatorname{Im} G = (E_{11}\mathscr{X}_1 + \operatorname{Im} G_{11}) \oplus (E_{22}\vec{\mathscr{C}}(E_{22}, F_{22}, G_{22}) + \operatorname{Im} G_{22})$.

Proof. Since the inclusions $\vec{\mathscr{C}}(E, F, G) \subset \mathscr{X}_1 \oplus \vec{\mathscr{C}}(E_{22}, F_{22}, G_{22})$ and $\mathscr{X}_1 \subset \vec{\mathscr{C}}(E, F, G)$ are obvious, it is sufficient to show that $\vec{\mathscr{C}}(E_{22}, F_{22}, G_{22}) \subset \vec{\mathscr{C}}(E, F, G)$. For this, let $\gamma := \max(\gamma(E_{11}, F_{11}, G_{11}), \gamma(E_{22}, F_{22}, G_{22}))$. Then, for each $x^2 \in \vec{\mathscr{C}}(E_{22}, F_{22}, G_{22})$, there exists $((x_k^2), (u_k^2)) \in \vec{\mathfrak{S}}(E_{22}, F_{22}, G_{22})$ such that $x_0^2 = 0$ and $x_\gamma^2 = x^2$. Since $\mathscr{X}_1 = \vec{\mathscr{C}}(E_{11}, F_{11}, G_{11})$ and the considered decomposition is good, it follows from Corollary 1.1(1) that the system $\vec{\mathfrak{S}}(E_{11}, F_{11}, G_{11})$ accepts all disturbance sequences from $\mathfrak{s}_0^+(\mathscr{X}_1)$. Moreover, in view of Lemma 1.1, $\lambda(E_{11}, F_{11}, G_{11}; \mathscr{X}_1) = 0$. Hence there exists $((\bar{x}_k^1), (\bar{u}_k^1)) \in \mathfrak{s}_0^+(\mathscr{X}_1) \times \mathfrak{s}_0^+(\mathscr{U}_1)$ such that $((\bar{x}_k^1), (\bar{u}_k^1), (-E_{12}x_{k+1}^2 + F_{12}x_k^2 + G_{12}u_k^2)) \in \vec{\mathfrak{S}}_d(E_{11}, F_{11}, G_{11})$ and $\bar{x}_0^1 = 0$. Since $\mathscr{X}_1 = \vec{\mathscr{C}}(E_{11}, F_{11}, G_{11})$ we can find $((\tilde{x}_k^1), (\tilde{u}_k^1)) \in \vec{\mathfrak{S}}(E_{11}, F_{11}, G_{11})$ satisfying $\tilde{x}_0^1 = 0$ and $\tilde{x}_\gamma^1 = \bar{x}_\gamma^1$. Now let $(x_k) := (\bar{x}_k^1) - (\tilde{x}_k^1) + (x_k^2)$ and $(u_k) := (\bar{u}_k^1) - (\tilde{u}_k^1) + (u_k^2)$. We can check that $((x_k), (u_k)) \in \vec{\mathfrak{S}}(E, F, G)$, $x_0 = 0$ and $x_\gamma = x^2$. Hence $\vec{\mathscr{C}}(E_{22}, F_{22}, G_{22}) \subset \vec{\mathscr{C}}(E, F, G)$. We have shown that $\vec{\mathscr{C}}(E, F, G) = \mathscr{X}_1 \oplus \vec{\mathscr{C}}(E_{22}, F_{22}, G_{22})$. To prove the remaining part of the theorem first note that $\mathscr{Z}_1 = E\mathscr{X}_1 + F\mathscr{X}_1 + G\mathscr{U}_1 = E_{11}\mathscr{X}_1 + F_{11}\mathscr{X}_1 + \operatorname{Im} G_{11} = E_{11}\mathscr{X}_1 + \operatorname{Im} G_{11}$; the last equality follows from Proposition 1.4(5). Hence

$$\begin{aligned} E\vec{\mathscr{C}}(E, F, G) + \operatorname{Im} G &= E\mathscr{X}_1 + E\vec{\mathscr{C}}(E_{22}, F_{22}, G_{22}) + \operatorname{Im} G \\ &= (E_{11}\mathscr{X}_1 + \operatorname{Im} G_{11}) \oplus (E_{22}\vec{\mathscr{C}}(E_{22}, F_{22}, G_{22}) + \operatorname{Im} G_{22}). \end{aligned}$$ □

In the present paper we shall frequently consider systems satisfying the condition $\mathscr{Z} = E\vec{\mathscr{C}}(E, F, G) + \operatorname{Im} G$. Such systems (which are called in [6], [8], and [9] *strongly controllable*) have various important properties. In particular, we can formulate the following result.

PROPOSITION 2.5. *Let* $(\mathscr{X}_1, \mathscr{X}_2; \mathscr{Z}_1, \mathscr{Z}_2; \mathscr{U})$ *be a decomposition of* $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$ *for the triple* (E, F, G) *such that the* 10*-tuple corresponding to* (E, F, G) *for this decomposition takes the form* $(E_{11}, E_{12}, 0, E_{22}; F_{11}, F_{12}, 0, F_{22}; G_1, G_2)$. *Assume that* $\mathscr{Z} = E\vec{\mathscr{C}}(E, F, G) + \operatorname{Im} G$. *Then* $\mathscr{X}_2 = \vec{\mathscr{C}}(E_{22}, F_{22}, G_2)$ *and* $\mathscr{Z}_2 = E_{22}\vec{\mathscr{C}}(E_{22}, F_{22}, G_2) + \operatorname{Im} G_2$.

Proof. In view of Propositions 1.4(5) and 1.2(6), $\mathscr{Z} = E\vec{\mathscr{C}}(E, F, G) + \operatorname{Im} G$ implies that $\vec{\mathscr{C}}(E, F, G) = \vec{V}(E, F, G)$. On the other hand, Proposition 1.2(4) yields $\vec{V}(E, F, G) = \mathscr{X}$. It ensures that $\mathscr{X} = \vec{\mathscr{C}}(E, F, G)$ and hence $\mathscr{X}_2 = \vec{\mathscr{C}}(E_{22}, F_{22}, G_2)$. Let us note also that $\mathscr{Z} = E\vec{\mathscr{C}}(E, F, G) + \operatorname{Im} G$ implies $\mathscr{Z}_2 = E_{22}\mathscr{X}_2 + \operatorname{Im} G_2$. □

Observe that in the terminology of [6] the proposition above states that $\vec{\mathfrak{S}}(E_{22}, F_{22}, G_2)$ is strongly controllable if $\vec{\mathfrak{S}}(E, F, G)$ enjoys the same property. Thus Proposition 2.5 generalizes in a sense [36, Prop. 1.2].

We end this section by recording the following simple result.

PROPOSITION 2.6. *Let* $(\mathscr{X}_1, \mathscr{X}_2; \mathscr{Z}_1, \mathscr{Z}_2; \mathscr{U})$ *be a decomposition of* $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$, *and let* $(E_{11}, E_{12}, 0, E_{22}; F_{11}, F_{12}, 0, F_{22}; G_1, G_2)$ *be the* 10*-tuple corresponding to* (E, F, G) *for this decomposition. Then* $\vec{\mathfrak{S}}_d(E, F, G)$ *possesses the uniqueness property if* $\vec{\mathfrak{S}}_d(E_{11}, F_{11}, G_1)$ *and* $\vec{\mathfrak{S}}_d(E_{22}, F_{22}, G_2)$ *enjoy the same property.*

3. Strongly almost invariant subspaces. Let us begin with the following definition.

DEFINITION 3.1. We say that a subspace $\mathcal{W} \subset \mathcal{X}$ is *strongly almost* $\vec{\mathcal{E}}(E, F, G)$-*invariant* if and only if it is almost $\vec{\mathcal{E}}(E, F, G)$-invariant and for any $((x_k), (u_k), (z_k)) \in \vec{\mathcal{E}}_d(E, F, G)$ such that $x_0 \in \mathcal{W}$ and $(z_k) \in \sigma_0^+(E\mathcal{W} + F\mathcal{W} + \operatorname{Im} G)$ the condition $(x_k) \in \sigma_0^+(\mathcal{W})$ is satisfied.

The following theorem provides a fundamental characterization of strongly almost invariant subspaces.

THEOREM 3.1. *Let* $(\mathcal{X}_1, \mathcal{X}_2; \mathcal{Z}_1, \mathcal{Z}_2; \mathcal{U})$ *be a good decomposition of* $(\mathcal{X}, \mathcal{Z}, \mathcal{U})$ *for the triple* (E, F, G). *Let* $(E_{11}, E_{12}, 0, E_{22}; F_{11}, F_{12}, 0, F_{22}; G_1, 0)$ *be the* 10*-tuple corresponding to the triple* (E, FG) *for this decomposition. Then the following statements are equivalent*:

(a) $\mathcal{X}_1$ *is strongly almost* $\vec{\mathcal{E}}(E, F, G)$*-invariant.*

(b) $\mathcal{X}_1$ *is almost* $\vec{\mathcal{E}}(E, F, G)$*-invariant and* $\vec{\mathcal{C}}(E, F, G) \subset \mathcal{X}_1$.

(c) $\vec{\mathcal{E}}_d(E_{11}, F_{11}, G_1)$ *accepts all disturbance sequences from* $\sigma_0^+(\mathcal{Z}_1)$ *and* $\vec{\mathcal{E}}_d(E_{22}, F_{22}, 0)$ *possesses the uniqueness property.*

Proof. (a)$\Rightarrow$(b) It is sufficient to show that $\vec{\mathcal{C}}(E, F, G) \subset \mathcal{X}_1$. For this consider an arbitrary $x \in \vec{\mathcal{C}}(E, F, G)$. It follows from Proposition 1.4(2) that there exist $i \in \mathbb{Z}_0^+$ and $((x_k), (u_k)) \in \vec{\mathcal{E}}(E, F, G)$ satisfying $x_0 = 0$ and $x_i = x$. Since $((x_k), (u_k), 0) \in \vec{\mathcal{E}}_d(E, F, G)$, $0 = x_0 \in \mathcal{X}_1$ and $\mathcal{X}_1$ is strongly almost $\vec{\mathcal{E}}(E, F, G)$-invariant, $(x_k) \in \sigma_0^+(\mathcal{X}_1)$ and hence $x \in \mathcal{X}_1$.

(b)$\Rightarrow$(c) In view of Proposition 2.1, $\vec{\mathcal{E}}_d(E_{11}, F_{11}, G_1)$ accepts all disturbance sequences from $\sigma_0^+(\mathcal{Z}_1)$, since $\mathcal{X}_1$ is assumed to be almost $\vec{\mathcal{E}}(E, F, G)$-invariant. To prove that $\vec{\mathcal{E}}_d(E_{22}, F_{22}, 0)$ possesses the uniqueness property, let us consider any sequence (x_k^2) such that $x_0^2 = 0$ and $((x_k^2), 0, 0) \in \vec{\mathcal{E}}_d(E_{22}, F_{22}, 0)$. Put $\lambda := \lambda(E_{11}, F_{11}, G_1; \mathcal{Z}_1)$ (cf. Proposition 1.5). Let $\bar{x}_{k+\lambda}^2 := x_k^2$, for $k \in \mathbb{Z}_0^+$, and $\bar{x}_k^2 := 0$, for $k = 0, 1, \cdots, \lambda - 1$. It is immediate that $((\bar{x}_k^2), 0, 0) \in \vec{\mathcal{E}}_d(E_{22}, F_{22}, 0)$. It follows from the equivalence (a)$\Leftrightarrow$(c) of Proposition 1.5 that there exists $((\bar{x}_k^1), (\bar{u}_k^1)) \in \sigma_0^+(\mathcal{X}) \times \sigma_0^+(\mathcal{U})$ such that $\bar{x}_0^1 = 0$ and $((\bar{x}_k^1), (\bar{u}_k^1), (-E_{12}\bar{x}_{k+1}^2 + F_{12}\bar{x}_k^2)) \in \vec{\mathcal{E}}_d(E_{11}, F_{11}, G_1)$. It is easily seen that $((\bar{x}_k^1 + \bar{x}_k^2), (\bar{u}_k^1)) \in \vec{\mathcal{E}}(E, F, G)$ and $\bar{x}_0^1 + \bar{x}_0^2 = 0$. Hence $(\bar{x}_k^1 + \bar{x}_k^2) \in \sigma_0^+(\vec{\mathcal{C}}(E, F, G))$ and, since $\vec{\mathcal{C}}(E, F, G) \subset \mathcal{X}_1$, $(\bar{x}_k^2) = 0$. But then, of course, $(x_k^2) = 0$.

(c)$\Rightarrow$(a) In view of Proposition 2.1, $\mathcal{X}_1$ is almost $\vec{\mathcal{E}}(E, F, G)$-invariant. Let $((x_k), (u_k), (z_k)) \in \vec{\mathcal{E}}_d(E, F, G)$ be such that $(z_k) \in \sigma_0^+(\mathcal{Z}_1)$ and $x_0 \in \mathcal{X}_1$. Then there exist (uniquely defined) $(x_k^i) \in \sigma_0^+(\mathcal{X}_i)$, for $i = 1, 2$, such that $(x_k) = (x_k^1) + (x_k^2)$. We can easily check that $((x_k^2), 0, 0) \in \vec{\mathcal{E}}_d(E_{22}, F_{22}, 0)$. But $x_0^2 = 0$; hence the uniqueness assumption ensures that $(x_k^2) = 0$, i.e., $(x_k) \in \sigma_0^+(\mathcal{X}_1)$. □

COROLLARY 3.1. *The space* $\vec{\mathcal{V}}(E, F, G) + \vec{\mathcal{R}}(E, F, G)$ *is the greatest strongly almost* $\vec{\mathcal{E}}(E, F, G)$*-invariant subspace. The space* $\vec{\mathcal{C}}(E, F, G)$ *is the smallest strongly almost* $\vec{\mathcal{E}}(E, F, G)$*-invariant subspace.*

Proof. The proof is an obvious consequence of Corollary 1.1, Proposition 1.2(6) and the equivalence (a)$\Leftrightarrow$(b) of Theorem 3.1. □

Another consequence of Theorem 3.1 is the following.

PROPOSITION 3.1. *Let* $(\mathcal{X}_1, \mathcal{X}_2; \mathcal{Z}_1, \mathcal{Z}_2; \mathcal{U})$ *be a Kalman decomposition of* $(\mathcal{X}, \mathcal{Z}, \mathcal{U})$ *for the triple* (E, F, G). *Let* $(E_{11}, E_{12}, 0, E_{22}; F_{11}, F_{12}, 0, F_{22}; G_1, 0)$ *be the* 10*-tuple corresponding to* (E, F, G) *for this decomposition. Let* $\mathcal{W}_2 := \mathcal{W} \cap \mathcal{X}_2$, *where* $\mathcal{W}$ *is a strongly almost* $\vec{\mathcal{E}}(E, F, G)$*-invariant subspace. Then* $\mathcal{W} = \mathcal{X}_1 \oplus \mathcal{W}_2$, $E\mathcal{W} + F\mathcal{W} + \operatorname{Im} G = \mathcal{Z}_1 \oplus (E_{22}\mathcal{W}_2 + F_{22}\mathcal{W}_2)$, *and* $\mathcal{W}_2$ *is almost* $\vec{\mathcal{E}}(E_{22}, F_{22}, 0)$*-invariant. Conversely, let* $\mathcal{W}_2 \subset \mathcal{X}_2$ *be any almost* $\vec{\mathcal{E}}(E_{22}, F_{22}, 0)$*-invariant subspace. Then* $\mathcal{X}_1 \oplus \mathcal{W}_2$ *is strongly almost* $\vec{\mathcal{E}}(E, F, G)$*-invariant.*

Proof. Since $\mathcal{X}_1 \subset \mathcal{W}$ (cf. Theorem 3.1) the equality $\mathcal{W} = \mathcal{X}_1 \oplus (\mathcal{W} \cap \mathcal{X}_2) = \mathcal{X}_1 \oplus \mathcal{W}_2$ can be checked using the modular distributive law. From this the equality $E\mathcal{W} + F\mathcal{W} +$

Im $G = \mathscr{X}_1 \oplus (E_{22}\mathscr{W}_2 + F_{22}\mathscr{W}_2)$ follows immediately. Note that $\mathscr{X}_1 = \vec{\mathscr{C}}(E, F, G)$ is almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant (cf. Corollary 1.1(1)). Hence Proposition 2.1 ensures that $\tilde{\mathfrak{S}}_d(E_{11}, F_{11}, G_1)$ accepts all disturbance sequences from $\mathscr{o}_0^+(\mathscr{X}_1)$. Now, the almost $\tilde{\mathfrak{S}}(E_{22}, F_{22}, 0)$-invariance of $\mathscr{W}_2$ follows from Proposition 2.2.

The proof of the second statement is an immediate consequence of Proposition 2.2 and Theorem 3.1. □

Before formulating our next result we recall (cf. [28], [30]) the following definition.

DEFINITION 3.2. A subspace $\mathscr{W} \subset \mathscr{X}$ is said to be an *(E&F)-deflating subspace* if and only if the condition $\dim \mathscr{W} = \dim (E\mathscr{W} + F\mathscr{W})$ holds.

PROPOSITION 3.2. *Let $\tilde{\mathfrak{S}}_d(E, F, 0)$ be a given regular system. Then the set of all almost $\tilde{\mathfrak{S}}(E, F, 0)$-invariant subspaces coincides with the set of all (E&F)-deflating subspaces. Moreover, the set is a sublattice of Lat $(\mathscr{X})$.*

Proof. See Corollary 4.1 and Theorem 4.2 of [5] for the proof. □

The theorem to be given below shows that some properties of strongly almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant subspaces can be expressed by corresponding properties of a certain (regular) subsystem of $\tilde{\mathfrak{S}}(E, F, G)$. This fact will allow us to use some results of [5].

THEOREM 3.2. *Let $(E, F, G) \in \mathscr{L}(\mathscr{X}, \mathscr{Z}) \times \mathscr{L}(\mathscr{X}, \mathscr{Z}) \times \mathscr{L}(\mathscr{U}, \mathscr{Z})$ be a given triple of linear maps. Let $(\mathscr{X}_1, \mathscr{X}_2, \mathscr{X}_3; \mathscr{Z}_1, \mathscr{Z}_2, \mathscr{Z}_3; \mathscr{U})$ be a Kalman decomposition of $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$ satisfying, in addition, $\mathscr{X}_1 \oplus \mathscr{X}_2 = \vec{\mathscr{V}}(E, F, G) + \vec{\mathscr{R}}(E, F, G)$ and $\mathscr{Z}_1 \oplus \mathscr{Z}_2 = E\vec{\mathscr{V}}(E, F, G) + F\vec{\mathscr{R}}(E, F, G) + \operatorname{Im} G$. Assume that $(E_{11}, \cdots, G_3)$ is the 21-tuple corresponding to (E, F, G) for this decomposition. Then the following statements hold.*

(1) $E_{21} = F_{21} = 0$, $E_{31} = F_{31} = 0$, $E_{32} = F_{32} = 0$, $G_2 = 0$, $G_3 = 0$.

(2) *The system $\tilde{\mathfrak{S}}_d(E_{22}, F_{22}, 0)$ is regular.*

(3) *The set of all almost $\tilde{\mathfrak{S}}(E_{22}, F_{22}, 0)$-invariant subspaces forms a sublattice of* Lat $(\mathscr{X}_2)$.

(4) *The set of all strongly almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant subspaces forms a sublattice of* Lat $(\mathscr{X})$. *Moreover, this lattice is isomorphic with the lattice of all almost $\tilde{\mathfrak{S}}(E_{22}, F_{22}, 0)$-invariant subspaces.*

Proof. We begin by observing that the decompositions $(\mathscr{X}_1, \mathscr{X}_2, \mathscr{X}_3; \mathscr{Z}_1, \mathscr{Z}_2, \mathscr{Z}_3; \mathscr{U})$ and $(\mathscr{X}_1 \oplus \mathscr{X}_2, \mathscr{X}_3; \mathscr{Z}_1 \oplus \mathscr{Z}_2, \mathscr{Z}_3; \mathscr{U})$ are good (cf. Corollary 1.1 and the equivalence (b)⇔(c) of Proposition 1.6).

(1) The proof of (1) is immediate in view of the above.

(2) Let $(\tilde{E}_{11}, \cdots, \tilde{G}_2)$ be the 10-tuple corresponding to the triple (E, F, G) for the decomposition $(\mathscr{X}_1 \oplus \mathscr{X}_2, \mathscr{X}_3; \mathscr{Z}_1 \oplus \mathscr{Z}_2, \mathscr{Z}_3; \mathscr{U})$. By Proposition 2.1, the system $\tilde{\mathfrak{S}}_d(\tilde{E}_{11}, \tilde{F}_{11}, \tilde{G}_1)$ accepts all disturbance sequences from $\mathscr{o}_0^+(\mathscr{Z}_1 \oplus \mathscr{Z}_2)$. Now applying Proposition 2.2 for the system $\tilde{\mathfrak{S}}_d(\tilde{E}_{11}, \tilde{F}_{11}, \tilde{G}_1)$, we obtain that the system $\tilde{\mathfrak{S}}_d(E_{22}, F_{22}, 0)$ accepts all disturbance sequences from $\mathscr{o}_0^+(\mathscr{Z}_2)$. Now it remains to prove that $\tilde{\mathfrak{S}}_d(E_{22}, F_{22}, 0)$ possesses the uniqueness property. For this, we first note that the implication (a)⇒(c) of Theorem 3.1 guarantees that $\tilde{\mathfrak{S}}_d(\tilde{E}_{22}, \tilde{F}_{22}, 0)$ possesses the uniqueness property. This fact allows us to prove that $\mathscr{X}_1 = \vec{\mathscr{C}}(E, F, G) = \vec{\mathscr{C}}(\tilde{E}_{11}, \tilde{F}_{11}, \tilde{G}_1)$. It follows that $\mathscr{X}_1$ is strongly almost $\tilde{\mathfrak{S}}(\tilde{E}_{11}, \tilde{F}_{11}, \tilde{G}_1)$-invariant. Now, again using the implication (a)⇒(c) of Theorem 3.1, we obtain that the system $\tilde{\mathfrak{S}}_d(E_{22}, F_{22}, 0)$ possesses the uniqueness property.

(3) The proof of part (3) follows from statement (2) and Proposition 3.2.

(4) We begin by noting that the map $\mathscr{W}_2 \mapsto \vec{\mathscr{C}}(E, F, G) \oplus \mathscr{W}_2$ from the Lat $(\mathscr{X}_2)$ into the Lat $(\mathscr{X})$ is a morphism of lattices, i.e., it preserves sums and intersections of subspaces. Using Proposition 3.1 and statement (3), we obtain that the set of all strongly almost $\tilde{\mathfrak{S}}(\tilde{E}_{11}, \tilde{F}_{11}, \tilde{G}_1)$-invariant subspaces forms a sublattice of Lat $(\mathscr{X}_1 \oplus \mathscr{X}_2)$. Let us now recall that $\mathscr{X}_1 \oplus \mathscr{X}_2 = \vec{\mathscr{V}}(E, F, G) + \vec{\mathscr{R}}(E, F, G)$ is the greatest strongly almost

$\tilde{\mathcal{E}}(E, F, G)$-invariant subspace of $\mathcal{X}$ (cf. Corollary 3.1). Hence, the set of all strongly almost $\tilde{\mathcal{E}}(\tilde{E}_{11}, \tilde{F}_{11}, \tilde{G}_1)$-invariant subspaces coincides with the set of all strongly almost $\tilde{\mathcal{E}}(E, F, G)$-invariant subspaces. □

The following generalization of Proposition 3.2 follows immediately from Theorems 4.1 and 4.2 of [5].

PROPOSITION 3.3. *Let $\tilde{\mathcal{E}}_d(E, F, D)$ be a given regular system. Then the set of all almost $\tilde{\mathcal{E}}(E, F, D)$-invariant subspaces $\mathcal{W}$ satisfying the inclusion* $\operatorname{Im} D \subset E\mathcal{W} + F\mathcal{W}$ *coincides with the set of all $(E\&F)$-deflating subspaces satisfying the same inclusion. Moreover, the set is a sublattice of* $\operatorname{Lat}(\mathcal{X})$. *The smallest element of this sublattice is $\vec{\mathcal{C}}(E, F, D)$.*

Now we are ready to establish the main result of this section.

THEOREM 3.3. *Let $(E, F, G) \in \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{U}, \mathcal{Z})$ be a given triple of linear maps, let $\mathcal{D}$ be a subspace of $\mathcal{Z}$, and let $D \in \mathcal{L}(\mathcal{D}, \mathcal{Z})$ be the canonical injection from $\mathcal{D}$ into $\mathcal{Z}$. Assume that $\tilde{\mathcal{E}}_d(E, F, G)$ accepts all disturbance sequences from $\mathcal{S}_0^+(\mathcal{D})$. Let $(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3; \mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3; \mathcal{U}, \mathcal{D})$ be a decomposition of $(\mathcal{X}, \mathcal{Z}, \mathcal{U} \times \mathcal{D})$ satisfying $\mathcal{X}_1 = \vec{\mathcal{C}}(E, F, G)$, $\mathcal{X}_1 \oplus \mathcal{X}_2 = \vec{\mathcal{V}}(E, F, G) + \vec{\mathcal{R}}(E, F, G)$, $\mathcal{Z}_1 = E\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G$, and $\mathcal{Z}_1 \oplus \mathcal{Z}_2 = E\vec{\mathcal{V}}(E, F, G) + F\vec{\mathcal{R}}(E, F, G) + \operatorname{Im} G$. Assume that $(E_{11}, \cdots, F_{33}; G_1, D_1, G_2, D_2, G_3, D_3)$ is the 24-tuple corresponding $(E, F, G \times D)$ for this $(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3; \mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3; \mathcal{U}, \mathcal{D})$. Then the following statements hold:*

(1) $E_{21} = F_{21} = 0$, $E_{31} = F_{31} = 0$, $E_{32} = F_{32} = 0$, $G_2 = 0$, $G_3 = 0$, $D_3 = 0$.

(2) *The set of all almost $\tilde{\mathcal{E}}(E_{22}, F_{22}, 0)$-invariant subspaces $\mathcal{W}_2$ satisfying* $\operatorname{Im} D_2 \subset E_{22}\mathcal{W}_2 + F_{22}\mathcal{W}_2$ *forms a sublattice of* $\operatorname{Lat}(\mathcal{X}_2)$. *The smallest element of this sublattice is $\vec{\mathcal{C}}(E_{22}, F_{22}, D_2)$.*

(3) $\vec{\mathcal{C}}(E, F, G \times D) = \vec{\mathcal{C}}(E, F, G) \oplus \vec{\mathcal{C}}(E_{22}, F_{22}, D)$ *and*

$$E\vec{\mathcal{C}}(E, F, G \times D) + F\vec{\mathcal{C}}(E, F, G \times D) + \operatorname{Im} G$$
$$= (E\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G) \oplus (E_{22}\vec{\mathcal{C}}(E_{22}, F_{22}, D_2) + F_{22}\vec{\mathcal{C}}(E_{22}, F_{22}, D_2)).$$

(4) *The set of all strongly almost $\tilde{\mathcal{E}}(E, F, G)$-invariant subspaces $\mathcal{W}$ satisfying $\mathcal{D} \subset E\mathcal{W} + F\mathcal{W} + \operatorname{Im} G$ forms a sublattice of* $\operatorname{Lat}(\mathcal{X})$. *The smallest element of this sublattice is $\vec{\mathcal{C}}(E, F, G \times D)$. Moreover, this sublattice is isomorphic to the lattice of all almost $\tilde{\mathcal{E}}(E_{22}, F_{22}, 0)$-invariant subspaces $\mathcal{W}_2$ satisfying* $\operatorname{Im} D_2 \subset E_{22}\mathcal{W}_2 + F_{22}\mathcal{W}_2$.

Proof. (1) Since the system $\tilde{\mathcal{E}}_d(E, F, G)$ accepts all disturbance sequences from $\mathcal{S}_0^+(\mathcal{D})$ the implication (a)⇒(b) of Proposition 1.5 yields that

$$\mathcal{D} \subset \mathcal{Z}_1 \oplus \mathcal{Z}_2 = E\vec{\mathcal{V}}(E, F, G) + F\vec{\mathcal{R}}(E, F, G) + \operatorname{Im} G.$$

Hence $D_3 = 0$. The rest of the proof follows from Theorem 3.2(1).

(2) The proof of part (2) follows from Theorem 3.2(2) and Proposition 3.3.

(3) Let $(\tilde{E}_{11}, \cdots, \tilde{F}_{22}; \tilde{G}_1, \tilde{D}_1, 0, 0)$ be the 12-tuple corresponding to the triple (E, F, G) for the decomposition $(\mathcal{X}_1 \oplus \mathcal{X}_2, \mathcal{X}_3; \mathcal{Z}_1 \oplus \mathcal{Z}_2, \mathcal{Z}_3; \mathcal{U}, \mathcal{D})$. By Corollary 3.1 $\mathcal{X}_1 \oplus \mathcal{X}_2$ is strongly almost $\tilde{\mathcal{E}}(E, F, G)$-invariant. Since the decomposition $(\mathcal{X}_1 \oplus \mathcal{X}_2, \mathcal{X}_3; \mathcal{Z}_1 \oplus \mathcal{Z}_2, \mathcal{Z}_3; \mathcal{U})$ is good for the triple (E, F, G) we can use the implication (a)⇒(c) of Theorem 3.1 to show that the system $\tilde{\mathcal{E}}_d(\tilde{E}_{22}, \tilde{F}_{22}, 0)$ possesses the uniqueness property. Hence $\vec{\mathcal{C}}(E, F, G \times D) = \vec{\mathcal{C}}(\tilde{E}_{11}, \tilde{F}_{11}, \tilde{G}_1 \times \tilde{D}_1)$. Now it remains to use Proposition 2.4 in the system $\tilde{\mathcal{E}}_d(\tilde{E}_{11}, \tilde{F}_{11}, \tilde{G}_1 \times \tilde{D}_1)$ to get

$$\vec{\mathcal{C}}(E, F, G \times D) = \vec{\mathcal{C}}(E, F, G) \oplus \vec{\mathcal{C}}(E_{22}, F_{22}, D_2)$$

and

$$E\vec{\mathcal{C}}(E, F, G \times D) + F\vec{\mathcal{C}}(E, F, G \times D) + \operatorname{Im} G = (E\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G)$$
$$\oplus (E_{22}\vec{\mathcal{C}}(E_{22}, F_{22}, D_2) + \operatorname{Im} D_2).$$

But, in view of Propositions 3.3 and 1.2(6),

$$E_{22}\vec{\mathcal{C}}(E_{22}, F_{22}, D_2) + \operatorname{Im} D_2 = E_{22}\vec{\mathcal{C}}(E_{22}, F_{22}, D_2) + F_{22}\vec{\mathcal{C}}(E_{22}, F_{22}, D_2).$$

(4) Part (4) can be proved using the same argument as in the proofs of Theorem 3.2(4) and Theorem 3.3(2), (3). □

Theorem 3.3 allows us to prove the following corollary.

COROLLARY 3.2. *Let* $(E, F, G) \in \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{U}, \mathcal{Z})$ *be a given triple of linear maps, let* $\mathcal{D}$ *be a subspace of* $\mathcal{Z}$, *and let* $D \in \mathcal{L}(\mathcal{D}, \mathcal{Z})$ *be the canonical injection from* $\mathcal{D}$ *into* $\mathcal{Z}$. *Assume that* $\vec{\mathfrak{S}}_d(E, F, G)$ *accepts all disturbance sequences from* $\mathcal{D}_0^+(\mathcal{D})$. *Then* $\mathcal{D} \subset E\vec{\mathcal{C}}(E, F, G \times D) + F\vec{\mathcal{C}}(E, F, G \times D) + \operatorname{Im} G$ *and*

$$\dim (E\vec{\mathcal{C}}(E, F, G \times D) + F\vec{\mathcal{C}}(E, F, G \times D) + \operatorname{Im} G) - \dim \vec{\mathcal{C}}(E, F, G \times D)$$
$$= \dim (E\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G) - \dim \vec{\mathcal{C}}(E, F, G).$$

In particular, when $G = 0$, $\mathcal{D} \subset E\vec{\mathcal{C}}(E, F, D) + F\vec{\mathcal{C}}(E, F, D)$ *and* $\dim (E\vec{\mathcal{C}}(E, F, D) + F\vec{\mathcal{C}}(E, F, D)) \leqq \dim \vec{\mathcal{C}}(E, F, D)$.

Proof. The first statement of the corollary is a consequence of Theorem 3.3(4). To prove the second statement, let us note that with the notation of Theorem 3.3(3) we have

$$\vec{\mathcal{C}}(E, F, G \times D) = \vec{\mathcal{C}}(E, F, G) \oplus \vec{\mathcal{C}}(E_{22}, F_{22}, D_2)$$

and

$$E\vec{\mathcal{C}}(E, F, G \times D) + F\vec{\mathcal{C}}(E, F, G \times D) + \operatorname{Im} G$$
$$= (E\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G) \oplus (E_{22}\vec{\mathcal{C}}(E_{22}, F_{22}, D_2) + F_{22}\vec{\mathcal{C}}(E_{22}, F_{22}, D_2)).$$

In view of Theorem 3.2(2) the system $\vec{\mathfrak{S}}_d(E_{22}, F_{22}, 0)$ is regular. But $\vec{\mathcal{C}}(E_{22}, F_{22}, D_2)$ is a deflating subspace in view of Proposition 3.3, thus

$$\dim (E_{22}\vec{\mathcal{C}}(E_{22}, F_{22}, D_2) + F_{22}\vec{\mathcal{C}}(E_{22}, F_{22}, D_2)) = \dim \vec{\mathcal{C}}(E_{22}, F_{22}, D_2).$$

The rest of the proof is a matter of an elementary calculation. □

4. Coasting and sliding subspaces. In this section we will generalize the concept of sliding and coasting subspaces introduced for a standard system by Willems [33]. This concept has been applied in [25] to study the problem of regularizability of the system. The coasting and sliding subspaces are defined there with the aid of subspace recursions and so their dynamical meaning is not obvious. In the present paper we give definitions by means of dynamical reasoning.

DEFINITION 4.1. A subspace $\mathcal{W} \subset \mathcal{X}$ will be called a *coasting subspace for* $\vec{\mathfrak{S}}(E, F, G)$ if and only if it is a uniqueness subspace for $\vec{\mathfrak{S}}(E, F, G)$, which is also $\vec{\mathfrak{S}}(E, F, G)$-invariant.

DEFINITION 4.2. A subspace $\mathcal{W} \subset \mathcal{X}$ is said to be a *sliding subspace for* $\vec{\mathfrak{S}}(E, F, G)$ if and only if it is simultaneously a uniqueness and a reachability subspace for $\vec{\mathfrak{S}}(E, F, G)$.

Note that $\mathcal{W}$ is a coasting subspace if and only if $\mathcal{W} = \vec{\mathcal{V}}(E, F, G)|\mathcal{W}$ and $\vec{\mathcal{C}}(E, F, G)|\mathcal{W} = 0$. Similarly, $\mathcal{W}$ is a sliding subspace if and only if $\mathcal{W} = \vec{\mathcal{R}}(E, F, G)|\mathcal{W}$ and $\vec{\mathcal{C}}(E, F, G)|\mathcal{W} = 0$.

We shall need in the sequel the following lemmas.

LEMMA 4.1. *Let* $\mathcal{R}$, $\mathcal{S}$, *and* $\mathcal{T}$ *be subspaces of* $\mathcal{X}$ *such that* $\mathcal{R} \subset \mathcal{S} + \mathcal{T}$ *and* $\mathcal{R} \cap \mathcal{T} = 0$. *Then there exists a subspace* $\mathcal{Q} \subset \mathcal{X}$ *satisfying* $\mathcal{Q} \subset \mathcal{S}$, $\mathcal{Q} \approx \mathcal{R}$, *and* $\mathcal{R} \oplus \mathcal{T} = \mathcal{Q} \oplus \mathcal{T}$.

LEMMA 4.2. *Let* $\mathcal{W} \subset \mathcal{X}$ *be any* $\vec{\mathfrak{S}}(E, F, G)$*-invariant subspace. Then for all* $j \in \mathbb{Z}_0^+$ *and for all* $i = 0, \cdots, j-1$ *there exists* $\mathcal{W}_i \subset \mathcal{X}$ *such that*

$$\mathcal{W} = \mathcal{W}_i \oplus \vec{\mathcal{C}}_j(E, F, G)|\mathcal{W} \quad \text{and} \quad \mathcal{W}_i \subset F^{-1}(E\mathcal{W}_i + \operatorname{Im} G) + \vec{\mathcal{C}}_{j-i-1}(E, F, G)|\mathcal{W}.$$

Proof. We first prove the statement for $i = 0$. For this let a subspace $\mathcal{W}_0$ be an arbitrary direct complement of $\vec{\mathcal{C}}_i(E, F, G)|\mathcal{W}$ to the space $\mathcal{W}$. Applying Propositions

1.2(1) and 1.4(5) (the latter for $\vec{\mathfrak{S}}(E, F, G)|\mathcal{W}$ and the modular distributive law we obtain

$$\begin{aligned}\mathcal{W}_0 \subset \mathcal{W} \subset F^{-1}(E\mathcal{W}+\operatorname{Im} G) &= F^{-1}(E\mathcal{W}_0+E\vec{\mathcal{C}}_j(E, F, G)|\mathcal{W}+\operatorname{Im} G)\\ &= F^{-1}(E\mathcal{W}_0+F\vec{\mathcal{C}}_{j-1}(E, F, G)|\mathcal{W}+\operatorname{Im} G)\\ &= F^{-1}(E\mathcal{W}_0+\operatorname{Im} G)+\vec{\mathcal{C}}_{j-1}(E, F, G)|\mathcal{W}.\end{aligned}$$

So, for $i=0$ the statement holds. Now, let the statement be true for some $i=0, \cdots, j-2$, i.e., $\mathcal{W}=\mathcal{W}_i\oplus\vec{\mathcal{C}}_j(E, F, G)|\mathcal{W}$ and $\mathcal{W}_i\subset F^{-1}(E\mathcal{W}_i+\operatorname{Im} G)+\vec{\mathcal{C}}_{j-i-1}(E, F, G)|\mathcal{W}$, for some $\mathcal{W}_i\subset\mathcal{X}$. By Proposition 1.4(1) $\vec{\mathcal{C}}_{j-i-1}(E, F, G)|\mathcal{W}\subset\vec{\mathcal{C}}_j(E, F, G)|\mathcal{W}$, so $\mathcal{W}_i\cap\vec{\mathcal{C}}_{j-i-1}(E, F, G)|\mathcal{W}=0$. In view of Lemma 4.1, there exists a subspace $\mathcal{W}_{i+1}\subset F^{-1}(E\mathcal{W}_i+\operatorname{Im} G)$ such that $\mathcal{W}_i\oplus\vec{\mathcal{C}}_{j-i-1}(E, F, G)|\mathcal{W}=\mathcal{W}_{i+1}\oplus\vec{\mathcal{C}}_{j-i-1}(E, F, G)|\mathcal{W}$. Hence we obtain $\mathcal{W}=\mathcal{W}_i\oplus\vec{\mathcal{C}}_j(E, F, G)|\mathcal{W}=\mathcal{W}_{i+1}+\vec{\mathcal{C}}_j(E, F, G)|\mathcal{W}$. But $\mathcal{W}_i\approx\mathcal{W}_{i+1}$ so $\mathcal{W}=\mathcal{W}_{i+1}\oplus\vec{\mathcal{C}}_j(E, F, G)|\mathcal{W}$. Now applying again Proposition 1.4(5) and the modular distributive law, we get

$$\begin{aligned}\mathcal{W}_{i+1}\subset F^{-1}(E\mathcal{W}_i+\operatorname{Im} G) &\subset F^{-1}(E\mathcal{W}_{i+1}+E\vec{\mathcal{C}}_{j-i-1}(E, F, G)|\mathcal{W}+\operatorname{Im} G)\\ &= F^{-1}(E\mathcal{W}_{i+1}+F\vec{\mathcal{C}}_{j-i-2}(E, F, G)|\mathcal{W}+\operatorname{Im} G)\\ &= F^{-1}(E\mathcal{W}_{i+1}+\operatorname{Im} G)+\vec{\mathcal{C}}_{j-i-2}(E, F, G)|\mathcal{W}.\end{aligned}$$

□

The theorem below is well known for a standard system (cf., e.g., [33, Thm. 7], [29, Lemma 2.24]). Let us recall that the "classical" proof of this theorem requires construction of a feedback map. In this context note that the proof presented below is purely geometric.

THEOREM 4.1. *Let $(E, F, G)\in\mathcal{L}(\mathcal{X}, \mathcal{Z})\times\mathcal{L}(\mathcal{X}, \mathcal{Z})\times\mathcal{L}(\mathcal{U}, \mathcal{Z})$ be given, and let $\mathcal{W}$ be any $\vec{\mathfrak{S}}(E, F, G)$-invariant subspace. Then there exists a coasting subspace $\mathcal{S}$ for the system $\vec{\mathfrak{S}}(E, F, G)$ such that $\mathcal{W}=\vec{\mathcal{C}}(E, F, G)|\mathcal{W}\oplus\mathcal{S}$ and $E\mathcal{W}+\operatorname{Im} G=(E\vec{\mathcal{C}}(E, F, G)|\mathcal{W}+\operatorname{Im} G)\oplus E\mathcal{S}$.*

Proof. It follows from Lemma 4.2 (for $j=\gamma(E, F, G)+1$ and $i=\gamma(E, F, G)$) that for every $\vec{\mathfrak{S}}(E, F, G)$-invariant subspace $\mathcal{W}$ there exists an $\vec{\mathfrak{S}}(E, F, G)$-invariant subspace $\mathcal{S}$ such that $\mathcal{W}=\vec{\mathcal{C}}(E, F, G)|\mathcal{W}\oplus\mathcal{S}$. Since $\vec{\mathcal{C}}(E, F, G)|\mathcal{S}\subset(\vec{\mathcal{C}}(E, FG)|\mathcal{W})\cap\mathcal{S}=0$, $\mathcal{S}$ is a coasting subspace of $\vec{\mathfrak{S}}(E, F, G)$. It is obvious that $E\mathcal{W}+\operatorname{Im} G=E\vec{\mathcal{C}}(E, F, G)|\mathcal{W}+E\mathcal{S}+\operatorname{Im} G$. Proposition 1.4(5), states that $E\vec{\mathcal{C}}(E, F, G)|\mathcal{W}+\operatorname{Im} G=F\vec{\mathcal{C}}(E, F, G)|\mathcal{W}+\operatorname{Im} G$. Hence to complete the proof we shall show that $(F\vec{\mathcal{C}}(E, F, G)|\mathcal{W}+\operatorname{Im} G)\cap E\mathcal{S}=0$. For this we consider arbitrary $z\in(F\vec{\mathcal{C}}(E, F, G)|\mathcal{W}+\operatorname{Im} G)\cap E\mathcal{S}$. Then we can find $c\in\vec{\mathcal{C}}(E, F, G)|\mathcal{W}$, $u\in\mathcal{U}$, and $s\in\mathcal{S}$ satisfying $z=Es=Fc+Gu$. Since $\mathcal{W}=\vec{\mathcal{V}}(E, F, G)|\mathcal{W}$ we can apply Proposition 1.2(6) to get the relation $s\in E^{-1}(F\vec{\mathcal{C}}(E, F, G)|\mathcal{W}+\operatorname{Im} G)\cap\mathcal{W}=\vec{\mathcal{C}}(E, F, G)|\mathcal{W}$. But $s\in\mathcal{S}$, so $s=0$, and consequently $z=0$. □

The following lemma is quite useful.

LEMMA 4.3. $\vec{\mathcal{V}}(E, F, G)=\vec{\mathcal{V}}(E, F, G)\cap\vec{\mathcal{V}}(E, F, G)+\vec{\mathcal{R}}(E, F, G)$.

Proof. The proof follows from Proposition 1.1 and Corollary 2.2 and Proposition 2.7 of [3]. □

Lemma 4.3 allows us to generalize Theorem 7 of [33]. Note again that the proof presented below uses an argument different from that of Theorem 7 of [33].

THEOREM 4.2. *Let $(E, F, G)\in\mathcal{L}(\mathcal{X}, \mathcal{Z})\times\mathcal{L}(\mathcal{X}, \mathcal{Z})\times\mathcal{L}(\mathcal{U}, \mathcal{Z})$ be given, and let $\mathcal{W}$ be any reachability subspace for the system $\vec{\mathfrak{S}}(E, F, G)$. Then there exists a sliding subspace $\mathcal{S}$ for the system $\vec{\mathfrak{S}}(E, F, G)$ such that $\mathcal{W}=\vec{\mathcal{C}}(E, F, G)|\mathcal{W}\oplus\mathcal{S}$ and $F\mathcal{W}+\operatorname{Im} G=(E\vec{\mathcal{C}}(E, F, G)|\mathcal{W}+\operatorname{Im} G)\oplus F\mathcal{S}$.*

Proof. Every reachability subspace $\mathcal{W}$ for $\vec{\mathfrak{S}}(E, F, G)$ is $\vec{\mathfrak{S}}(E, F, G)$-invariant. Therefore (cf. Proposition 1.2(2)) $\mathcal{W}$ is also $\vec{\mathfrak{S}}(F, E, G)$-invariant. Applying Theorem

4.1 to the system $\vec{\Sigma}(F, E, G)$, we obtain that there exists a coasting subspace $\mathcal{S}$ for $\vec{\Sigma}(F, E, G)$ such that $\mathcal{W} = \vec{\mathcal{C}}(F, E, G)|\mathcal{W} \oplus \mathcal{S}$ and $F\mathcal{W} + \operatorname{Im} G = (F\vec{\mathcal{C}}(F, E, G)|\mathcal{W} + \operatorname{Im} G) \oplus F\mathcal{S}$. Applying Proposition 1.4(4), (5), we obtain $F\vec{\mathcal{C}}(F, E, G)|\mathcal{W} + \operatorname{Im} G = E\vec{\mathcal{C}}(E, F, G)|\mathcal{W} + \operatorname{Im} G$. So it remains to prove that $\mathcal{S}$ is a reachability subspace for $\vec{\Sigma}(E, F, G)$. Since $\mathcal{S} = \vec{\mathcal{V}}(E, F, G)|\mathcal{S}$, then by Lemma 4.3 we get the equality $\mathcal{S} = (\vec{\mathcal{V}}(E, F, G)|\mathcal{S}) \cap (\vec{\mathcal{V}}(E, F, G)|\mathcal{S}) + \vec{\mathcal{R}}(E, F, G)|\mathcal{S}$. Note, that since $\mathcal{W}$ is a reachability subspace for $\vec{\Sigma}(E, F, G)$, $\vec{\mathcal{V}}(E, F, G)|\mathcal{W} = (\vec{\mathcal{V}}(E, F, G)|\mathcal{W}) \cap (\vec{\mathcal{R}}(E, F, G)|\mathcal{W}) = \vec{\mathcal{C}}(E, F, G)|\mathcal{W}$. Now applying Proposition 1.4(4), we obtain $\vec{\mathcal{V}}(E, F, G)|\mathcal{W} = \vec{\mathcal{C}}(F, E, G)|\mathcal{W}$. It yields $(\vec{\mathcal{V}}(E, F, G)|\mathcal{S}) \cap (\vec{\mathcal{V}}(E, F, G)|\mathcal{S}) \subset \mathcal{S} \cap (\vec{\mathcal{V}}(E, F, G)|\mathcal{W}) = \mathcal{S} \cap (\vec{\mathcal{C}}(F, E, G)|\mathcal{W})$. But $\mathcal{S}$ is a coasting subspace $\mathcal{S}$ for $\vec{\Sigma}(F, E, G)$. Hence $\mathcal{S} \cap (\vec{\mathcal{C}}(F, E, G)|\mathcal{W}) = 0$. We have shown that $\mathcal{S} = \vec{\mathcal{R}}(E, F, G)|\mathcal{S}$, so $\mathcal{S}$ is a reachability subspace for $\vec{\Sigma}(E, F, G)$. Since $\vec{\mathcal{C}}(E, F, G)|\mathcal{S} = \vec{\mathcal{C}}(F, E, G)|\mathcal{S} = 0$, $\mathcal{S}$ is a sliding subspace for $\vec{\Sigma}(E, F, G)$. □

The corollary below summarizes the results of Theorems 4.1 and 4.2. For a standard control system it reduces to Theorem 7 of [33] or Theorem 2.27 of [29]. If $\mathcal{W} = \mathcal{X}$, a similar result is announced in Theorem 3.1 of [25] for a class of implicit systems.

COROLLARY 4.1. *Let $\mathcal{W}$ be any almost $\vec{\Sigma}(E, F, G)$-invariant subspace. Then there exist a coasting subspace $\mathcal{T}$ and a sliding subspace $\mathcal{S}$ for the system $\vec{\Sigma}(E, F, G)$ such that $\mathcal{W} = \vec{\mathcal{C}}(E, F, G)|\mathcal{W} \oplus \mathcal{T} \oplus \mathcal{S}$ and $E\mathcal{W} + F\mathcal{W} + \operatorname{Im} G = (E\vec{\mathcal{C}}(E, F, G)|\mathcal{W} + \operatorname{Im} G) \oplus E\mathcal{T} \oplus F\mathcal{S}$. In particular, when $\mathcal{W}$ is strongly almost $\vec{\Sigma}(E, F, G)$-invariant, $\mathcal{W} = \vec{\mathcal{C}}(E, F, G) \oplus \mathcal{T} \oplus \mathcal{S}$ and $E\mathcal{W} + F\mathcal{W} + \operatorname{Im} G = (E\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G) \oplus E\mathcal{T} \oplus F\mathcal{S}$.*

Proof. First observe that for arbitrary $\mathcal{W} \subset \mathcal{X}$, $\vec{\mathcal{V}}(E, F, G)|\mathcal{W}$ is an $\vec{\Sigma}(E, F, G)$-invariant subspace, while $\vec{\mathcal{R}}(E, F, G)|\mathcal{W}$ is a reachability subspace for $\vec{\Sigma}(E, F, G)$. So we can use Theorems 4.1 and 4.2 to obtain $\vec{\mathcal{V}}(E, F, G)|\mathcal{W} = \vec{\mathcal{C}}(E, F, G)|\mathcal{W} \oplus \mathcal{T}$ and $\vec{\mathcal{R}}(E, F, G)|\mathcal{W} = \vec{\mathcal{C}}(E, F, G)|\mathcal{W} \oplus \mathcal{S}$, for some coasting subspace $\mathcal{T}$ and sliding subspace $\mathcal{S}$. Note that $\mathcal{T} \cap \vec{\mathcal{R}}(E, F, G)|\mathcal{W} \subset \vec{\mathcal{V}}(E, F, G)|\mathcal{W} \cap \vec{\mathcal{R}}(E, F, G)|\mathcal{W} = \vec{\mathcal{C}}(E, F, G)|\mathcal{W}$, hence $\mathcal{T} \cap \vec{\mathcal{R}}(E, F, G)|\mathcal{W} \subset \mathcal{T} \cap \vec{\mathcal{C}}(E, F, G)|\mathcal{W} = 0$. Using a similar argument, we can prove that $\mathcal{S} \cap \vec{\mathcal{V}}(E, F, G)|\mathcal{W} = 0$ and $\vec{\mathcal{C}}(E, F, G)|\mathcal{W} \cap (\mathcal{T} \oplus \mathcal{S}) = 0$. By Proposition 1.6 $\mathcal{W} = \vec{\mathcal{V}}(E, F, G)|\mathcal{W} + \vec{\mathcal{R}}(E, F, G)|\mathcal{W}$; thus $\mathcal{W} = \vec{\mathcal{C}}(E, F, G)|\mathcal{W} \oplus \mathcal{T} \oplus \mathcal{S}$.

To prove the second equality consider an arbitrary

$$z \in (E\vec{\mathcal{V}}(E, F, G)|\mathcal{W} + \operatorname{Im} G) \cap F\mathcal{S} = (E\vec{\mathcal{C}}(E, F, G)|\mathcal{W} + E\mathcal{T} + \operatorname{Im} G) \cap F\mathcal{S}.$$

Then there exist $v \in \vec{\mathcal{V}}(E, F, G)|\mathcal{W}$, $u \in \mathcal{U}$, and $s \in \mathcal{S}$ such that $z = Ev + Gu = Fs$. Hence, by Proposition 1.2(4),

$$\begin{aligned} s &\in \mathcal{S} \cap (F^{-1}(E\vec{\mathcal{V}}(E, F, G)|\mathcal{W} + \operatorname{Im} G)) \cap \mathcal{W} \\ &= \mathcal{S} \cap \vec{\mathcal{V}}(E, F, G)|\mathcal{W} \subset \mathcal{S} \cap \vec{\mathcal{C}}(E, F, G)|\mathcal{W} = 0. \end{aligned}$$

So $z = 0$ and therefore $(E\vec{\mathcal{C}}(E, F, G)|\mathcal{W} + E\mathcal{T} + \operatorname{Im} G) \cap F\mathcal{S} = 0$. By a similar argument, we obtain $(E\vec{\mathcal{C}}(E, F, G)|\mathcal{W} + F\mathcal{S} + \operatorname{Im} G) \cap E\mathcal{T} = 0$ and $(E\vec{\mathcal{C}}(E, F, G)|\mathcal{W} + \operatorname{Im} G) \cap (E\mathcal{T} \oplus F\mathcal{S}) = 0$. Since (by Proposition 1.6)

$$\begin{aligned} E\mathcal{W} + F\mathcal{W} + \operatorname{Im} G &= E\vec{\mathcal{V}}(E, F, G)|\mathcal{W} + F\vec{\mathcal{R}}(E, F, G)|\mathcal{W} + \operatorname{Im} G \\ &= E\vec{\mathcal{C}}(E, F, G)|\mathcal{W} + E\mathcal{T} + F\mathcal{S} + \operatorname{Im} G \end{aligned}$$

we get the desired equality.

The rest of the proof follows immediately from the equivalence (a)⇔(b) of Theorem 3.1. □

Remark 4.1. Note that the proof of Lemma 4.2 is constructive. Therefore we can compute explicitly coasting and sliding subspaces with the properties described by Theorems 4.1 and 4.2 and Corollary 4.1.

We also record the following simple result.

PROPOSITION 4.1. *A subspace $\mathcal{T} \subset \mathcal{X}$ is a coasting subspace for $\tilde{\mathfrak{S}}(E, F, G)$ if and only if it is $\tilde{\mathfrak{S}}(E, F, G)$-invariant and $\mathcal{T} \cap \vec{\mathcal{C}}_1(E, F, G) = 0$. In particular, $E\mathcal{T} \approx \mathcal{T}$ if $\mathcal{T}$ is a coasting subspace for $\tilde{\mathfrak{S}}(E, F, G)$. Similarly, a subspace $\mathcal{S} \subset \mathcal{X}$ is a sliding subspace for $\tilde{\mathfrak{S}}(E, F, G)$ if and only if it is a reachability subspace for $\tilde{\mathfrak{S}}(E, F, G)$ and $\mathcal{S} \cap \vec{\mathcal{C}}_1(F, E, G) = 0$. In particular, $F\mathcal{S} \approx \mathcal{S}$ if $\mathcal{S}$ is a sliding subspace for $\tilde{\mathfrak{S}}(E, F, G)$.*

Proof. Note that for any $\tilde{\mathfrak{S}}(E, F, G)$-invariant subspace $\mathcal{T}$ $\vec{\mathcal{C}}_1(E, F, G)|\mathcal{T} = \mathcal{T} \cap \vec{\mathcal{C}}_1(E, F, G)$. Since, by Proposition 1.4(3) $\vec{\mathcal{C}}(E, F, G)|\mathcal{T}$ is the smallest $\tilde{\mathfrak{S}}(E, F, G)$-invariant subspace containing $\vec{\mathcal{C}}_1(E, F, G)|\mathcal{T}$, $\vec{\mathcal{C}}(E, F, G)|\mathcal{T} = 0$ if and only if $\vec{\mathcal{C}}_1(E, F, G)|\mathcal{T} = 0$. We have shown that $\mathcal{T}$ is a coasting subspace if and only if it is $\tilde{\mathfrak{S}}(E, F, G)$-invariant and $\mathcal{T} \cap \vec{\mathcal{C}}_1(E, F, G) = 0$. To prove that $E\mathcal{T} \approx \mathcal{T}$ for a sliding subspace $\mathcal{T}$, it is sufficient to note that $\operatorname{Ker} E \cap \mathcal{W} \subset \vec{\mathcal{C}}_1(E, F, G) \cap \mathcal{W}$ for any $\mathcal{W}$ being $\tilde{\mathfrak{S}}(E, F, G)$-invariant.

The proof of the second statement is similar to the first and so is omitted. □

In the following definition we distinguish a certain class of almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant subspaces.

DEFINITION 4.3. A subspace $\mathcal{W} \subset \mathcal{X}$ is called a *regularizing subspace for the system* $\tilde{\mathfrak{S}}(E, F, G)$ if and only if it is almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant and it is a uniqueness subspace for $\tilde{\mathfrak{S}}(E, F, G)$.

In other words, $\mathcal{W}$ is a regularizing subspace for the system $\tilde{\mathfrak{S}}(E, F, G)$ if and only if $\mathcal{W}$ is almost $\tilde{\mathfrak{S}}(E, F, G)$-invariant and $\vec{\mathcal{C}}(E, F, G)|\mathcal{W} = 0$. It happens that regularizing subspaces play an important role in solving the DDPU.

We shall need the following equivalent characterization of regularizing subspaces for a system $\tilde{\mathfrak{S}}(E, F, 0)$.

PROPOSITION 4.2. *Let $(E, F, 0) \in \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{U}, \mathcal{Z})$. Assume $(\mathcal{X}_1, \mathcal{X}_2; \mathcal{Z}_1, \mathcal{Z}_2; \mathcal{U})$ to be a good decomposition of $(\mathcal{X}; \mathcal{Z}; \mathcal{U})$, and $(E_{11}, \cdots, F_{22}; 0, 0)$ to be the 10-tuple corresponding to the triple $(E, F, 0)$ for this decomposition. Then the following statements are equivalent:*

(a) *$\mathcal{X}_1$ is a regularizing subspace for $\tilde{\mathfrak{S}}(E, F, 0)$.*

(b) *$\tilde{\mathfrak{S}}_d(E_{11}, F_{11}, 0)$ is regular.*

(c) *$\mathcal{X}_1$ is an $(E\&F)$-deflating and a uniqueness subspace for $\tilde{\mathfrak{S}}(E, F, 0)$.*

(d) *There exist a coasting subspace $\mathcal{T}$ and a sliding subspace $\mathcal{S}$ for the system $\tilde{\mathfrak{S}}_d(E, F, 0)$ such that $\mathcal{X}_1 = \mathcal{T} \oplus \mathcal{S}$ and $\mathcal{Z}_1 = E\mathcal{T} \oplus F\mathcal{S}$.*

Proof. The proof follows easily from Propositions 1.7, 2.1, 4.1, and Corollary 4.1. □

PROPOSITION 4.3. *Let $(E, F) \in \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{X}, \mathcal{Z})$, $\mathcal{D}$ be a subspace of $\mathcal{Z}$, and let $D \in \mathcal{L}(\mathcal{D}, \mathcal{Z})$ be the canonical injection from $\mathcal{D}$ into $\mathcal{Z}$. Suppose that the system $\tilde{\mathfrak{S}}_d(E, F, 0)$ accepts all disturbance sequences from $\mathcal{S}_0^+(\mathcal{D})$. Then there exists a regularizing subspace $\mathcal{W}$ for the system $\tilde{\mathfrak{S}}_d(E, F, 0)$ such that $\mathcal{W} \subset \vec{\mathcal{C}}(E, F, D)$ and $\mathcal{D} \subset E\mathcal{W} + F\mathcal{W}$. Moreover, such a subspace $\mathcal{W}$ can be chosen so that if $(\mathcal{X}_1, \mathcal{X}_2; \mathcal{Z}_1, \mathcal{Z}_2; \mathcal{D})$ is a decomposition of $(\mathcal{X}; \mathcal{Z}; \mathcal{D})$ satisfying $\mathcal{X}_1 = \mathcal{W}$, $\mathcal{Z}_1 = E\mathcal{W} + F\mathcal{W}$, and $(E_{11}, \cdots, F_{22}; D_1, D_2)$ is the 10-tuple corresponding to (E, F, D) for this decomposition, then $E_{21} = F_{21} = 0$, $D_2 = 0$, the system $\tilde{\mathfrak{S}}_d(E_{11}, F_{11}, 0)$ is regular, and $\mathcal{X}_1 = \vec{\mathcal{C}}(E_{11}, F_{11}, D_1)$.*

Proof. Theorem 3.3(4) ensures that $\vec{\mathcal{C}}(E, F, D)$ is strongly almost $\tilde{\mathfrak{S}}(E, F, 0)$-invariant. Now, by Corollary 4.1, there exist a coasting subspace $\bar{\mathcal{T}}$ and a sliding subspace $\mathcal{S}$ for the system $\tilde{\mathfrak{S}}(E, F, 0)$ such that $\vec{\mathcal{C}}(E, F, D) = \vec{\mathcal{C}}(E, F, 0) \oplus \bar{\mathcal{T}} \oplus \mathcal{S}$ and $E\vec{\mathcal{C}}(E, F, D) + F\vec{\mathcal{C}}(E, F, D) = E\vec{\mathcal{C}}(E, F, 0) \oplus E\bar{\mathcal{T}} \oplus F\mathcal{S}$. Moreover, $\mathcal{D} \subset E\vec{\mathcal{C}}(E, F, 0) \oplus E\bar{\mathcal{T}} \oplus F\mathcal{S}$. Let $\bar{\mathcal{C}} \subset \mathcal{X}$ be such that $\bar{\mathcal{C}} \oplus (\vec{\mathcal{C}}(E, F, 0) \cap \operatorname{Ker} E) = \vec{\mathcal{C}}(E, F, 0)$. Put $\mathcal{T} := \bar{\mathcal{C}} \oplus \bar{\mathcal{T}}$. Then $\mathcal{T}$ is $\tilde{\mathfrak{S}}(E, F, 0)$-invariant by Propositions 1.3 and 1.2(3). Since $E\bar{\mathcal{C}} \approx \bar{\mathcal{C}}$, $E\bar{\mathcal{T}} \approx \bar{\mathcal{T}}$, and $E\mathcal{T} = E\bar{\mathcal{C}} \oplus E\bar{\mathcal{T}}$, we get $E\mathcal{T} \approx \mathcal{T}$. Now, by Proposition 4.1, $\mathcal{T}$ is a coasting subspace for the system $\tilde{\mathfrak{S}}(E, F, 0)$. Thus $\mathcal{D} \subset E\mathcal{T} \oplus F\mathcal{S}$. Let $(\bar{\mathcal{X}}_1, \bar{\mathcal{X}}_2; \bar{\mathcal{Z}}_1, \bar{\mathcal{Z}}_2; \mathcal{D})$ be a

decomposition of $(\mathscr{X}; \mathscr{Z}; \mathscr{D})$ such that $\bar{\mathscr{X}}_1 := \mathscr{T} \oplus \mathscr{S}$ and $\bar{\mathscr{Z}}_1 := E\mathscr{T} \oplus F\mathscr{S}$. Let $(\bar{E}_{11}, \cdots, \bar{F}_{22}; \bar{D}_1, \bar{D}_2)$ be the 10-tuple corresponding to (E, F, D) for this decomposition. Then $\bar{E}_{21} = \bar{F}_{21} = 0$ and $\bar{D}_2 = 0$. Note that by the implication (d)$\Rightarrow$(a) of Proposition 4.2, $\bar{\mathscr{X}}_1$ is a regularizing subspace. On the other hand, $\bar{\mathscr{Z}}_1 = E\bar{\mathscr{X}}_1 + F\bar{\mathscr{X}}_1$, in view of Proposition 1.2(1), (2). Now, by the implication (a)$\Rightarrow$(b) of Proposition 4.2, $\tilde{\mathfrak{S}}_d(\bar{E}_{11}, \bar{F}_{11}, 0)$ is regular. Let $\mathscr{W} := \tilde{\mathscr{C}}(\bar{E}_{11}, \bar{F}_{11}, \bar{D}_1)$. Proposition 3.3 ensures that $\mathscr{W}$ is an $(\bar{E}_{11}\&\bar{F}_{11})$-deflating subspace and $\operatorname{Im} \bar{D}_1 \subset \bar{E}_{11}\mathscr{W} + \bar{F}_{11}\mathscr{W}$. Since the considered decomposition of $(\mathscr{X}, \mathscr{Z}, \mathscr{U})$ is good, $\mathscr{W}$ is also $(E\&F)$-deflating and $\operatorname{Im} D \subset E\mathscr{W} + F\mathscr{W}$. The rest of the proof follows from Proposition 4.2. □

5. The disturbance decoupling problem. This section is essentially based on the results reported in [2].

We begin by reformulating the definition of the DDP and the DDPU given in § 0.

PROPOSITION 5.1. *Let*

$$(E, F, G, D, H) \in \mathscr{L}(\mathscr{X}, \mathscr{Z}) \times \mathscr{L}(\mathscr{X}, \mathscr{Z}) \times \mathscr{L}(\mathscr{U}, \mathscr{Z}) \times \mathscr{L}(\mathscr{D}, \mathscr{Z}) \times \mathscr{L}(\mathscr{X}, \mathscr{Y})$$

be a given quintuple of linear maps.

(1) *The* DDP *is solvable for the quintuple* (E, F, G, D, H) *if and only if there exists* $K \in \mathscr{L}(\mathscr{X}, \mathscr{U})$ *such that the system* $\tilde{\mathfrak{S}}_d(E, F + GK, 0)|\operatorname{Ker} H$ *accepts all disturbance sequences from* $\mathscr{I}_0^+(\operatorname{Im} D)$.

(2) *The* DDPU *is solvable for the quintuple* (E, F, G, D, H) *if and only if there exists* $K \in \mathscr{L}(\mathscr{X}, \mathscr{U})$ *such that the system* $\tilde{\mathfrak{S}}_d(E, F + GK, 0)|\operatorname{Ker} H$ *accepts all disturbance sequences from* $\mathscr{I}_0^+(\operatorname{Im} D)$ *and the system* $\tilde{\mathfrak{S}}_d(E, F + GK, G)$ *possesses the uniqueness property.*

It happens that for solving the above-stated problems it is of crucial importance to find necessary and sufficient conditions for a given subspace $\mathscr{W}$ to be $\tilde{\mathfrak{S}}(E, F + GK, 0)$-invariant for some $K \in \mathscr{L}(\mathscr{X}, \mathscr{U})$. These conditions have recently been found by Fletcher and Aasaraai in Theorem 2.1 of [17]. The conditions are $E\mathscr{W} \subset F\mathscr{W} + \operatorname{Im} G$ and $\dim(E\mathscr{W} \cap \operatorname{Im} G) \leqq \dim((F^{-1} \operatorname{Im} G) \cap \mathscr{W})$. In fact, for solving the DDP we should answer a more general question. This is done by the following lemma, the proof of which is a slight modification of that given in [17].

LEMMA 5.1. *Let* $\mathscr{W} \subset \mathscr{X}$ *and* $\mathscr{D} \subset \mathscr{Z}$. *Then there exists* $K \in \mathscr{L}(\mathscr{X}, \mathscr{U})$ *such that* $E\mathscr{W} + \mathscr{D} \subset (F + GK)\mathscr{W}$ *if and only if* $E\mathscr{W} + \mathscr{D} \subset F\mathscr{W} + \operatorname{Im} G$ *and* $\dim((E\mathscr{W} + \mathscr{D}) \cap \operatorname{Im} G) \leqq \dim((F^{-1} \operatorname{Im} G) \cap \mathscr{W})$.

Proof. ($\Rightarrow$) Let $(e_i)_{i=1}^k$ be a basis in $(E\mathscr{W} + \mathscr{D}) \cap \operatorname{Im} G$. Then there exists a linearly independent sequence $(w_i)_{i=1}^k$ such that $e_i = (F + GK)w_i$, $i = 1, \cdots, k$. Note that $w_i \in (F^{-1} \operatorname{Im} G) \cap \mathscr{W}$, since $e_i \in \operatorname{Im} G$, for $i = 1, \cdots, k$. Hence the subspace $F^{-1} \operatorname{Im} G \cap \mathscr{W}$ contains at least $k = \dim((E\mathscr{W} + \mathscr{D}) \cap \operatorname{Im} G)$ linearly independent vectors.

($\Leftarrow$) Let $(e_i)_{i=1}^d$ be a basis in $E\mathscr{W} + \mathscr{D}$ satisfying $\operatorname{span}\{e_i; i = 1, \cdots, k\} = (E\mathscr{W} + \mathscr{D}) \cap \operatorname{Im} G$. Then there exists $(\tilde{u}_i)_{i=1}^d$ and $(w_i)_{i=k+1}^d$ such that $e_i = G\tilde{u}_i$, for $i = 1, \cdots, k$, and $e_i = Fw_i + G\tilde{u}_i$, for $i = k+1, \cdots, d$. It is easy to check that $(\tilde{u}_i)_{i=1}^k$ are linearly independent and so are $(w_i)_{i=k+1}^d$. Since $k = \dim((E\mathscr{W} + \mathscr{D}) \cap \operatorname{Im} G) \leqq \dim((F^{-1} \operatorname{Im} G) \cap \mathscr{W})$ there exists a linearly independent sequence $(w_i)_{i=1}^k$ such that $w_i \in (F^{-1} \operatorname{Im} G) \cap \mathscr{W}$ for $i = 1, \cdots, k$. Hence we can find $(\bar{u}_i)_{i=1}^k$ such that $Fw_i = -G\bar{u}_i$, $i = 1, \cdots, k$. Let $u_i := \tilde{u}_i + \bar{u}_i$, for $i = 1, \cdots, k$ and $u_i := \tilde{u}_i$, for $i = k+1, \cdots, d$. Note that $e_i = Fw_i + Gu_i$, for $i = 1, \cdots, d$. To show that $(w_i)_{i=1}^d$ are linearly independent suppose $\sum_{i=1}^d \alpha_i w_i = 0$, for some real α_i, $i = 1, \cdots, d$. It yields that $e := \sum_{i=1}^d \alpha_i e_i = G\sum_{i=1}^d \alpha_i u_i \in (E\mathscr{W} + \mathscr{D}) \cap \operatorname{Im} G$, hence $\alpha_i = 0$, for $i = k+1, \cdots, d$. Since the sequence $(w_i)_{i=1}^k$ is linearly independent, we obtain $\alpha_i = 0$, for $i = 1, \cdots, k$. Now it is sufficient to choose any map $K \in \mathscr{L}(\mathscr{X}, \mathscr{U})$ satisfying $Kw_i = u_i$, $i = 1, \cdots, d$. □

Using Lemma 5.1, we can prove the following useful result.

LEMMA 5.2. *Let $\mathcal{W} \subset \mathcal{X}$ and $\mathcal{D} \subset \mathcal{X}$. Then there exists $K \in \mathcal{L}(\mathcal{X}, \mathcal{U})$ such that $E\mathcal{W} + \mathcal{D} \subset (F + GK)\mathcal{W}$ if $E\mathcal{W} + \mathcal{D} + \operatorname{Im} G = F\mathcal{W} + \operatorname{Im} G$ and $\dim (E\mathcal{W} + \mathcal{D}) \leqq \dim \mathcal{W}$.*

Proof. Let $\tilde{F} := F|\mathcal{W}$. Then it is easy to check that $(F^{-1} \operatorname{Im} G) \cap \mathcal{W} = \tilde{F}^{-1} \operatorname{Im} G$. Hence

$$\begin{aligned}\dim ((F^{-1} \operatorname{Im} G) \cap \mathcal{W}) &= \dim \tilde{F}^{-1} \operatorname{Im} G = \dim (\operatorname{Im} \tilde{F} \cap \operatorname{Im} G) + \dim \operatorname{Ker} \tilde{F} \\ &= \dim \operatorname{Im} \tilde{F} + \dim \operatorname{Im} G - \dim (\operatorname{Im} \tilde{F} + \operatorname{Im} G) + \dim \operatorname{Ker} \tilde{F} \\ &= \dim \mathcal{W} + \dim \operatorname{Im} G - \dim (F\mathcal{W} + \operatorname{Im} G).\end{aligned}$$

On the other hand,

$$\dim ((E\mathcal{W} + \mathcal{D}) \cap \operatorname{Im} G) = \dim (E\mathcal{W} + \mathcal{D}) + \dim \operatorname{Im} G - \dim (E\mathcal{W} + \mathcal{D} + \operatorname{Im} G).$$

Therefore $\dim ((E\mathcal{W} + \mathcal{D}) \cap \operatorname{Im} G) \leqq \dim ((F^{-1} \operatorname{Im} G) \cap \mathcal{W})$ if $E\mathcal{W} + \mathcal{D} + \operatorname{Im} G = F\mathcal{W} + \operatorname{Im} G$ and $\dim (E\mathcal{W} + \mathcal{D}) \leqq \dim \mathcal{W}$. Now the proof follows from Lemma 5.1. □

The main result of the paper is the following.

THEOREM 5.1. *Let $(E, F, G) \in \mathcal{L}(\mathcal{X}, \mathcal{X}) \times \mathcal{L}(\mathcal{X}, \mathcal{X}) \times \mathcal{L}(\mathcal{U}, \mathcal{X})$ be a given triple of linear maps, let $\mathcal{D}$ be a subspace of $\mathcal{X}$, and let $D \in \mathcal{L}(\mathcal{D}, \mathcal{X})$ be the canonical injection from $\mathcal{D}$ into $\mathcal{X}$. Then there exists $K \in \mathcal{L}(\mathcal{X}, \mathcal{U})$ such that $\vec{\mathfrak{S}}_d(E, F + GK, 0)$ accepts all disturbance from $\mathcal{S}_0^+(\mathcal{D})$ if and only if $\vec{\mathfrak{S}}_d(E, F, G)$ accepts all disturbance sequences from $\mathcal{S}_0^+(\mathcal{D})$ and $\dim (E\vec{\mathcal{C}}(E, F, G \times D) + \mathcal{D}) \leqq \dim \vec{\mathcal{C}}(E, F, G \times D)$.*

Proof. ($\Rightarrow$) We first prove that $\dim (E\vec{\mathcal{C}}(E, F, G \times D) + \mathcal{D}) \leqq \dim \vec{\mathcal{C}}(E, F, G \times D)$. Applying Corollary 3.2, we immediately obtain the inclusion $\mathcal{D} \subset E\vec{\mathcal{C}}(E, F + GK, D) + (F + GK)\vec{\mathcal{C}}(E, F + GK, D)$ and the inequality

$$\dim (E\vec{\mathcal{C}}(E, F + GK, D) + (F + GK)\vec{\mathcal{C}}(E, F + GK, D)) \leqq \dim \vec{\mathcal{C}}(E, F + GK, D).$$

Hence $\dim (E\vec{\mathcal{C}}(E, F + GK, D) + \mathcal{D}) \leqq \dim \vec{\mathcal{C}}(E, F + GK, D)$. Let us observe also that $\vec{\mathcal{C}}(E, F + GK, D) \subset \vec{\mathcal{C}}(E, F, G \times D)$. So we can find $\mathcal{W} \subset \mathcal{X}$ such that $\mathcal{W} \oplus \vec{\mathcal{C}}(E, F + GK, D) = \vec{\mathcal{C}}(E, F, G \times D)$. Therefore

$$\begin{aligned}\dim (E\vec{\mathcal{C}}(E, F, G \times D) + \mathcal{D}) &= \dim (E\mathcal{W} + E\vec{\mathcal{C}}(E, F + GK, D) + \mathcal{D}) \\ &\leqq \dim E\mathcal{W} + \dim (E\vec{\mathcal{C}}(E, F + GK, D) + \mathcal{D}) \\ &\leqq \dim \mathcal{W} + \dim \vec{\mathcal{C}}(E, F + GK, D) \\ &= \dim \vec{\mathcal{C}}(E, F, G \times D).\end{aligned}$$

To end the proof of the implication it is sufficient to note that the system $\vec{\mathfrak{S}}_d(E, F, G)$ accepts all disturbance from $\mathcal{S}_0^+(\mathcal{D})$ if $\vec{\mathfrak{S}}_d(E, F + GK, 0)$ accepts all disturbance sequences from $\mathcal{S}_0^+(\mathcal{D})$.

($\Leftarrow$) We begin by observing that, in view of Theorem 3.3(4) the subspace $\vec{\mathcal{C}}(E, F, G \times D)$ is strongly almost $\vec{\mathfrak{S}}_d(E, F, G)$-invariant. So we can use Corollary 4.1 to conclude that

$$\vec{\mathcal{C}}(E, F, G \times D) = \vec{\mathcal{C}}(E, F, G) \oplus \mathcal{T} \oplus \mathcal{S}$$

and

$$E\vec{\mathcal{C}}(E, F, G \times D) + F\vec{\mathcal{C}}(E, F, G \times D) + \operatorname{Im} G = (E\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G) \oplus E\mathcal{T} \oplus F\mathcal{S},$$

for some coasting subspace $\mathcal{T}$ and sliding subspace $\mathcal{S}$. Let $\mathcal{W} \subset \mathcal{X}$ be chosen so that $\vec{\mathcal{C}}(E, F, G) = \mathcal{W} \oplus (\vec{\mathcal{C}}(E, F, G) \cap \operatorname{Ker} E)$. Note that $\mathcal{W}$ is $\vec{\mathfrak{S}}(E, F, G)$-invariant and $E\mathcal{W} \approx \mathcal{W}$ (cf. Proposition 1.3). Let $(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3; \mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3; \mathcal{U}, \mathcal{D})$ be a decomposition of $(\mathcal{X}, \mathcal{Z}, \mathcal{U} \times \mathcal{D})$ satisfying $\mathcal{X}_1 = \mathcal{W} \oplus \mathcal{T}$, $\mathcal{X}_2 = (\vec{\mathcal{C}}(E, F, G) \cap \operatorname{Ker} E) \oplus \mathcal{S}$, $\mathcal{Z}_1 = E\mathcal{X}_1 = E\mathcal{W} \oplus E\mathcal{T}$, and

$$\begin{aligned}\mathcal{X}_1 \oplus \mathcal{X}_2 &= E\vec{\mathcal{C}}(E, F, G \times D) + F\vec{\mathcal{C}}(E, F, G \times D) + \operatorname{Im} G \\ &= (E\vec{\mathcal{C}}(E, F, G) + \operatorname{Im} G) \oplus E\mathcal{T} \oplus F\mathcal{S}.\end{aligned}$$

Assume that $(E_{11}, \cdots, F_{33}; G_1, D_1, G_2, D_2, G_3, D_3)$ is the 24-tuple corresponding to the triple $(E, F, G \times D)$ for the decomposition. Since $(\mathscr{X}_1 \oplus \mathscr{X}_2, \mathscr{X}_3; \mathscr{Z}_1 \oplus \mathscr{Z}_2, \mathscr{Z}_3; \mathscr{U}, \mathscr{D})$ is a good decomposition of $(\mathscr{X}, \mathscr{Z}, \mathscr{U} \times \mathscr{D})$ for the triple $(E, F, G \times D)$, $E_{31} = F_{31} = 0$, $E_{32} = F_{32} = 0$, $G_3 = 0$, and $D_3 = 0$. Similarly, $E\mathscr{X}_1 = \mathscr{Z}_1$ implies $E_{21} = 0$. Observe now that

$$\mathscr{Z}_1 \oplus \mathscr{Z}_2 = (E\vec{\mathscr{C}}(E, F, G) + \operatorname{Im} G) \oplus E\mathscr{T} \oplus F\mathscr{S} = E\mathscr{W} + E\mathscr{T} + \operatorname{Im} G + F\mathscr{S}$$

$$= \mathscr{Z}_1 + F\mathscr{S} + \operatorname{Im} G.$$

Hence $\mathscr{Z}_2 = F_{22}\mathscr{X}_2 + \operatorname{Im} G$. Note also that $E\vec{\mathscr{C}}(E, F, G \times D) + \mathscr{D} = E\mathscr{W} + E\mathscr{T} + E\mathscr{S} + \mathscr{D} = \mathscr{Z}_1 + E\mathscr{S} + \mathscr{D}$. Therefore $E\vec{\mathscr{C}}(E, F, G \times D) + \mathscr{D} = \mathscr{Z}_1 \oplus (E_{22}\mathscr{X}_2 + \operatorname{Im} D_2)$ and hence

$$\dim (E\vec{\mathscr{C}}(E, F, G \times D) + \mathscr{D}) = \dim \mathscr{Z}_1 + \dim (E_{22}\mathscr{X}_2 + \operatorname{Im} D_2).$$

On the other hand,

$$\dim \vec{\mathscr{C}}(E, F, G \times D) = \dim (\mathscr{X}_1 \oplus \mathscr{X}_2) = \dim \mathscr{X}_1 + \dim \mathscr{X}_2.$$

But

$$\dim \mathscr{Z}_1 = \dim (E\mathscr{W} \oplus E\mathscr{T}) = \dim E\mathscr{W} + \dim E\mathscr{T} = \dim \mathscr{W} + \dim \mathscr{T} = \dim \mathscr{X}_1,$$

since $E\mathscr{W} \approx \mathscr{W}$ and $E\mathscr{T} \approx \mathscr{T}$ (cf. Proposition 4.1). Thus $\dim (E\vec{\mathscr{C}}(E, F, G \times D) + \mathscr{D}) \leqq \dim \vec{\mathscr{C}}(E, F, G \times D)$ implies that $\dim (E_{22}\mathscr{X}_2 + \operatorname{Im} D_2) \leqq \dim \mathscr{X}_2$. We shall now prove that $E_{22}\mathscr{X}_2 + \operatorname{Im} G_2 + \operatorname{Im} D_2 = F_{22}\mathscr{X}_2 + \operatorname{Im} G_2$. First observe that (by Proposition 1.4(5))

$$F\vec{\mathscr{C}}(E, F, G \times D) + \operatorname{Im} G \subset E\vec{\mathscr{C}}(E, F, G \times D) + \operatorname{Im} G + \mathscr{D}.$$

It follows that $\mathscr{Z}_2 = F_{22}\mathscr{X}_2 + \operatorname{Im} G_2 \subset E_{22}\mathscr{X}_2 + \operatorname{Im} G_2 + \operatorname{Im} D_2$. Hence $\mathscr{Z}_2 = E_{22}\mathscr{X}_2 + \operatorname{Im} G_2 + \operatorname{Im} D_2 = F_{22}\mathscr{X}_2 + \operatorname{Im} G_2$. Now we can use Lemma 5.2 to show that there exists $K_2 \in \mathscr{L}(\mathscr{X}_2, \mathscr{U})$ such that $E_{22}\mathscr{X}_2 + \operatorname{Im} D_2 \subset (F_{22} + G_2K_2)\mathscr{X}_2$ and consequently $\vec{\mathscr{E}}_d(E_{22}, F_{22} + G_2K_2, 0)$ accepts all disturbance sequences from $\mathscr{I}_0^+(\operatorname{Im} D_2)$. Note that $\mathscr{X}_1$ being a sum of $\vec{\mathscr{E}}(E, F, G)$-invariant subspaces, it is also $\vec{\mathscr{E}}(E, F, G)$-invariant (cf. Proposition 1.2(3)). Therefore $F_{21}\mathscr{X}_1 \subset \operatorname{Im} G_2$ and hence we can find $K_1 \in \mathscr{L}(\mathscr{X}_1, \mathscr{U})$ such that $F_{21} + G_2K_1 = 0$. Let $K \in \mathscr{L}(\mathscr{X}, \mathscr{U})$ be such that $K|\mathscr{X}_1 = K_1$, $K|\mathscr{X}_2 = K_2$, and $K|\mathscr{X}_3 = 0$. Then the 24-tuple corresponding to the triple $(E, F + GK, G \times D)$ for the decomposition $(\mathscr{X}_1, \mathscr{X}_2, \mathscr{X}_3; \mathscr{Z}_1, \mathscr{Z}_2, \mathscr{Z}_3; \mathscr{U}, \mathscr{D})$ of $(\mathscr{X}, \mathscr{Z}, \mathscr{U} \times \mathscr{D})$ takes the form

$$(E_{11}, E_{12}, E_{13}, 0, E_{22}, E_{23}, 0, 0, E_{33}; F_{11} + G_1K_1, F_{12}$$

$$+ G_1K_2, F_{13}, 0, F_{22} + G_2K_2, F_{32}, 0, 0, F_{33}; G_1, D_1, G_2, D_2, 0, 0).$$

The map E_{11} is invertible so obviously $\vec{\mathscr{E}}_d(E_{11}, F_{11} + G_1K_1, 0)$ accepts all disturbance sequences from $\mathscr{I}_0^+(\mathscr{Z}_1)$. On the other hand, we know from the previous considerations that $\vec{\mathscr{E}}_d(E_{22}, F_{22} + G_2K_2, 0)$ accepts all disturbance sequences from $\mathscr{I}_0^+(\operatorname{Im} D_2)$. Hence, in view of Proposition 2.2, the system $\vec{\mathscr{E}}_d(E, F + GK, 0)$ accepts all disturbance sequences from $\mathscr{I}_0^+(\mathscr{D})$. □

Theorem 5.2 to be given below completely solves the DDP.

THEOREM 5.2. *Let*

$$(E, F, G, D, H) \in \mathscr{L}(\mathscr{X}, \mathscr{Z}) \times \mathscr{L}(\mathscr{X}, \mathscr{Z}) \times \mathscr{L}(\mathscr{U}, \mathscr{Z}) \times \mathscr{L}(\mathscr{D}, \mathscr{Z}) \times \mathscr{L}(\mathscr{X}, \mathscr{Y})$$

be a given quintuple of linear maps. Then the DDP *is solvable for the quintuple* (E, F, G, D, H) *if and only if the following conditions hold*:

(i) $\operatorname{Im} D \subset E\vec{\mathscr{V}}(E, F, G)|\operatorname{Ker} H + F\vec{\mathscr{R}}(E, F, G)|\operatorname{Ker} H + \operatorname{Im} G$;

(ii) $\dim (E\vec{\mathscr{C}}(E, F, G \times D)|\operatorname{Ker} H + \operatorname{Im} D) \leqq \dim \vec{\mathscr{C}}(E, F, G \times D)|\operatorname{Ker} H$.

Let us recall the concept of U-regularizability (cf. [9], [16], [25], [26]).

DEFINITION 5.1. We say that a system $\vec{\mathscr{E}}_d(E, F, G)$ is *U-regularizable* if and only if there exists $K \in \mathscr{L}(\mathscr{X}, \mathscr{U})$ such that the system $\vec{\mathscr{E}}_d(E, F + GK, G)$ possesses the uniqueness property.

To prove our next result we shall need the following lemma (cf. [9, Thm. 3.1]). (The proof of the result given in [9] is based on construction of a feedback map K such that $\operatorname{Ker}(F+GK)\cap\vec{\mathscr{C}}(E,F+GK,G)=0$.)

LEMMA 5.3. *Let* $(E,F,G)\in\mathscr{L}(\mathscr{X},\mathscr{Z})\times\mathscr{L}(\mathscr{X},\mathscr{Z})\times\mathscr{L}(\mathscr{U},\mathscr{Z})$ *be given. Then the system* $\vec{\mathfrak{S}}_d(E,F,G)$ *is U-regularizable if and only if* $\dim\vec{\mathscr{C}}(E,F,G)\leqq\dim(E\vec{\mathscr{C}}(E,F,G)+\operatorname{Im}G)$.

We can now prove the following theorem solving the DDPU.

THEOREM 5.3. *Let*

$$(E,F,G,D,H)\in\mathscr{L}(\mathscr{X},\mathscr{Z})\times\mathscr{L}(\mathscr{X},\mathscr{Z})\times\mathscr{L}(\mathscr{U},\mathscr{Z})\times\mathscr{L}(\mathscr{D},\mathscr{Z})\times\mathscr{L}(\mathscr{X},\mathscr{Y})$$

be a given quintuple of linear maps. Then the DDPU *is solvable for the quintuple* (E,F,G,D,H) *if and only if the following conditions hold*:

(i) $\operatorname{Im}D\subset E\vec{\mathscr{V}}(E,F,G)|\operatorname{Ker}H+F\vec{\mathscr{R}}(E,F,G)|\operatorname{Ker}H+\operatorname{Im}G$;

(ii) $\dim(E\vec{\mathscr{C}}(E,F,G\times D)|\operatorname{Ker}H+\operatorname{Im}D)\leqq\dim\vec{\mathscr{C}}(E,F,G\times D)|\operatorname{Ker}H$;

(iii) $\dim\vec{\mathscr{C}}(E,F,G)\leqq\dim(E\vec{\mathscr{C}}(E,F,G)+\operatorname{Im}G)$.

Proof. ($\Rightarrow$) The proof is obvious in view of Theorem 5.2 and Lemma 5.3.

($\Leftarrow$) We know from Theorem 5.2 that there exists $\bar{K}\in\mathscr{L}(\mathscr{X},\mathscr{U})$ such that $\vec{\mathfrak{S}}_d(E,F+G\bar{K},0)|\operatorname{Ker}H$ accepts all disturbance sequences from $\mathscr{d}_0^+(\operatorname{Im}D)$. In particular, $\vec{\mathfrak{S}}_d(E,F+G\bar{K},0)$ has the same property. Put $\bar{F}:=F+G\bar{K}$. By Proposition 4.3, there exists a regularizing subspace $\mathscr{X}_1\subset\mathscr{X}$ for the system $\vec{\mathfrak{S}}(E,\bar{F},0)$ such that $\mathscr{X}_1\subset\vec{\mathscr{C}}(E,\bar{F},D)$ and $\operatorname{Im}D\subset\mathscr{Z}_1:=E\mathscr{X}_1+\bar{F}\mathscr{X}_1$. It is easy to check that $\vec{\mathscr{C}}(E,\bar{F},D)\subset\vec{\mathscr{C}}(E,F,G\times D)$, so we obtain $\mathscr{X}_1\subset\vec{\mathscr{C}}(E,F,G\times D)$ and $\mathscr{Z}_1\subset E\vec{\mathscr{C}}(E,F,G\times D)+F\vec{\mathscr{C}}(E,F,G\times D)+\operatorname{Im}G$. Let $\mathscr{X}_2,\mathscr{X}_3,\mathscr{Z}_2$, and $\mathscr{Z}_3$ be such that $\mathscr{X}_1\oplus\mathscr{X}_2=\vec{\mathscr{C}}(E,F,G\times D)$, $\mathscr{X}_1\oplus\mathscr{X}_2\oplus\mathscr{X}_3=\mathscr{X}$, $\mathscr{Z}_1\oplus\mathscr{Z}_2=E\vec{\mathscr{C}}(E,F,G\times D)+F\vec{\mathscr{C}}(E,F,G\times D)+\operatorname{Im}G$, and $\mathscr{Z}_1\oplus\mathscr{Z}_2\oplus\mathscr{Z}_3=\mathscr{Z}$. Then $(\mathscr{X}_1,\mathscr{X}_2,\mathscr{X}_3;\mathscr{Z}_1,\mathscr{Z}_2,\mathscr{Z}_3;\mathscr{U},\mathscr{D})$ is a decomposition of $(\mathscr{X},\mathscr{Z},\mathscr{U}\times\mathscr{D})$. Assume that $(E_{11},\cdots,F_{33};G_1,D_1,G_2,D_2,G_3,D_3)$ is the 24-tuple corresponding to the triple $(E,F,G\times D)$ for this decomposition. Let $\bar{K}_i:=\bar{K}|\mathscr{X}_i$ and $\bar{F}_{ji}:=F_{ji}+G_j\bar{K}_i$, for $i,j=1,2,3$. It follows from the proof of Theorem 5.1 that we can choose $\bar{K}$ so that $\bar{K}_3=0$. Then $E_{21}=\bar{F}_{21}=F_{21}+G_2\bar{K}_1=0$, $E_{31}=\bar{F}_{31}=F_{31}=0$, $E_{32}=\bar{F}_{32}=F_{32}=0$, $D_2=0$, $G_3=0$, and $D_3=0$. Since, in view of Theorem 3.3(4), $\mathscr{X}_1\oplus\mathscr{X}_2=\vec{\mathscr{C}}(E,F,G\times D)$ is strongly almost $\vec{\mathfrak{S}}(E,F,G)$-invariant, we can use the equivalence (a)$\Leftrightarrow$(c) of Theorem 3.1 to conclude that $\vec{\mathfrak{S}}_d(E_{33},\bar{F}_{33},0)$ possesses the uniqueness property. Proposition 4.3 ensures that $\vec{\mathfrak{S}}_d(E_{11},\bar{F}_{11},0)$ is regular; hence $\dim\mathscr{X}_1=\dim\mathscr{Z}_1$. Thus

$$\begin{aligned}\dim(E\vec{\mathscr{C}}(E,F,G\times D)+F\vec{\mathscr{C}}(E,F,G\times D)+\operatorname{Im}G)-\dim\vec{\mathscr{C}}(E,F,G\times D)\\ =\dim\mathscr{Z}_2-\dim\mathscr{X}_2.\end{aligned}$$

Let us observe now that $\vec{\mathfrak{S}}(E,F,G\times D)$-invariance of $\vec{\mathscr{C}}(E,F,G\times D)$ implies that $\mathscr{Z}_1\oplus\mathscr{Z}_2=E\vec{\mathscr{C}}(E,F,G\times D)+\operatorname{Im}G+\operatorname{Im}D$. Hence Proposition 2.5 ensures that the equalities $\mathscr{Z}_2=E_{22}\vec{\mathscr{C}}(E_{22},\bar{F}_{22},G_2)+\operatorname{Im}G_2$ and $\mathscr{X}_2=\vec{\mathscr{C}}(E_{22},\bar{F}_{22},G_2)$ hold. Applying Corollary 3.2, we obtain

$$\dim\mathscr{Z}_2-\dim\mathscr{X}_2=\dim(E\vec{\mathscr{C}}(E,F,G)+\operatorname{Im}G)-\dim\vec{\mathscr{C}}(E,F,G).$$

This equality together with the inequality

$$\dim(E\vec{\mathscr{C}}(E,F,G)+\operatorname{Im}G)\geqq\dim\vec{\mathscr{C}}(E,F,G)$$

ensures us that

$$\dim\mathscr{Z}_2=\dim(E_{22}\vec{\mathscr{C}}(E_{22},\bar{F}_{22},G_2)+\operatorname{Im}G_2)\geqq\dim\mathscr{X}_2=\dim\vec{\mathscr{C}}(E_{22},\bar{F}_{22},G_2).$$

So using Lemma 5.3, we can find $\tilde{K}_2\in\mathscr{L}(\mathscr{X}_2,\mathscr{U})$ such that $\vec{\mathfrak{S}}_d(E_{22},\bar{F}_{22}+G_2\tilde{K}_2,0)$ possesses the uniqueness property. Let $\tilde{K}\in\mathscr{L}(\mathscr{X},\mathscr{U})$ satisfy $\tilde{K}|\mathscr{X}_1=0$, $\tilde{K}|\mathscr{X}_2=\tilde{K}_2$, and $\tilde{K}|\mathscr{X}_3=0$. Put $K:=\bar{K}+\tilde{K}$. Now the 24-tuple corresponding to the triple $(E,F+GK,D)$

takes the form

$$(E_{11}, E_{12}, E_{13}, 0, E_{22}, E_{23}, 0, 0, E_{33}; F_{11}+G_1\bar{K}_1, F_{12} + G_1(\bar{K}_2+\tilde{K}_2), F_{13}, 0, F_{22}+G_2(\bar{K}_2+\tilde{K}_2), F_{23}, 0, 0, F_{33}; D_1, 0, 0).$$

Since the systems $\tilde{\mathfrak{S}}_d(E_{11}, F_{11}+G_1\bar{K}_1, 0)$, $\tilde{\mathfrak{S}}_d(E_{22}, F_{22}+G_2(\bar{K}_2+\tilde{K}_2), 0)$, and $\tilde{\mathfrak{S}}_d(E_{33}, F_{33}, 0)$ possess the uniqueness property by Proposition 2.6 the same property has the system $\tilde{\mathfrak{S}}_d(E, F+GK, 0)$. Note also that since $\tilde{\mathfrak{S}}_d(E_{11}, F_{11}+G_1\bar{K}_1, 0)$ accepts all disturbance sequences from $\mathscr{V}_0^+(\operatorname{Im} D_1)$, the system $\tilde{\mathfrak{S}}_d(E, F+GK, 0)$ accepts all disturbance sequences from $\mathscr{V}_0^+(\operatorname{Im} D)$. □

COROLLARY 5.1. *Let* $(E, F, G, D, H) \in \mathscr{L}(\mathscr{X}, \mathscr{X}) \times \mathscr{L}(\mathscr{X}, \mathscr{X}) \times \mathscr{L}(\mathscr{U}, \mathscr{X}) \times \mathscr{L}(\mathscr{D}, \mathscr{X}) \times \mathscr{L}(\mathscr{X}, \mathscr{Y})$ *be a given quintuple of linear maps. Then the* DDPU *is solvable for the quintuple* (E, F, G, D, H) *if and only if the* DDP *is solvable for this quintuple and the system* $\tilde{\mathfrak{S}}_d(E, F, G)$ *is U-regularizable.*

COROLLARY 5.2. *Let* $\tilde{\mathfrak{S}}(E, F, G)$ *be a regular system. Then the* DDP *is solvable if and only if the* DDPU *is solvable.*

Remark 5.1. Note that the conditions of Theorems 5.2 and 5.3 are easily checked because the subspaces occurring in their formulation can be computed with the aid of some subspace recursions (cf. the Appendix). Since the coasting and sliding subspaces used in the proof of Theorem 5.1 can be determined explicitly (cf. Remark 4.1), we can compute a feedback map solving the DDP. The most important step for computing such a feedback is described in the (constructive) proof of Lemma 5.1. Let us now discuss the DDPU. To solve the problem we should first compute a preliminary feedback solving the DDP. Then it is sufficient to modify this feedback (as it is described in the proof Theorem 5.3) so that the obtained "closed-loop" system possesses the uniqueness property. But this can be made using standard methods of regularizability (cf. [9], [16], [25], [26]).

Example. Consider the DDPU for a quintuple (E, F, G, D, H) defined as follows:

$$E = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad F = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

$$G = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \qquad H = [0 \quad 0 \quad 0 \quad 0 \quad 1].$$

Let, for $i = 1, 2, \cdots, 5$, e_i denote the ith versor of $\mathbb{R}^5$. Since $\operatorname{Ker} H = \operatorname{span}\{e_1, e_2, e_3, e_4\}$, we obtain

$$E|\operatorname{Ker} H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad F|\operatorname{Ker} H = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

We can easily check (cf. Remark 1.1) that

$$\vec{\mathcal{V}}(E, F, G)|\operatorname{Ker} H = \operatorname{span}\{e_1, e_2\}, \qquad \vec{\mathcal{R}}(E, F, G)|\operatorname{Ker} H = \operatorname{span}\{e_1, e_3, e_4\},$$

$$\vec{\mathcal{C}}(E, F, G)|\operatorname{Ker} H = \operatorname{span}\{e_1\},$$

and

$$\vec{\mathcal{C}}(E, F, G \times D)|\operatorname{Ker} H = \operatorname{span}\{e_1, e_2, e_3, e_4\}.$$

Thus conditions (i)-(iii) of Theorem 5.3 hold and the DDPU is solvable. Let us now calculate the corresponding feedback map K. For this, decompose (according to Corollary 4.1) the subspace $\vec{\mathcal{C}}(E, F, G \times D)|\operatorname{Ker} H$ into a direct sum of $\vec{\mathcal{C}}(E, F, G)|\operatorname{Ker} H$, a coasting subspace $\mathcal{T}$, and a sliding subspace $\mathcal{S}$. Let us observe that it is sufficient to put $\mathcal{T} = \operatorname{span}\{e_2\}$ and $\mathcal{S} = \operatorname{span}\{e_3, e_4\}$. The construction of the feedback map now follows from Theorem 5.1 applied for the system $\tilde{\mathfrak{S}}_d(E, F, G)|\operatorname{Ker} H$. We can check that the map

$$K = \begin{bmatrix} k_{11} & k_{12} & k_{13} & k_{14} & k_{15} \\ -1 & 0 & -1 & -1 & k_{25} \end{bmatrix},$$

where $k_{11}, \cdots, k_{15}$ and k_{25} are arbitrary, solves the DDPU.

Let us note that here $E\vec{\mathcal{V}}(E, F, G)|\operatorname{Ker} H + F\vec{\mathcal{R}}(E, F, G)|\operatorname{Ker} H + \operatorname{Im} G = \mathcal{X}$. Hence, condition (i) of Theorem 5.2 (or equivalently of Theorem 5.3) is satisfied for each map D. The equivalence (a)$\Leftrightarrow$(c) of Proposition 1.5 ensures that the system $\tilde{\mathfrak{S}}_d(E, F, G)|\operatorname{Ker} H$ accepts all disturbance sequences from $\mathcal{A}_0^+(\operatorname{Im} D)$, independently of the map D. Thus the "open-loop" disturbance decoupling problem considered in [5] is solvable for every D. However we can easily find a map D for which condition (ii) of Theorem 5.2 does not hold and hence the DDP is not solvable for the quintuple (E, F, G, D, H). An example of such a map D reads as follows:

$$D = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Remark 5.2. In the second version of the paper [17] (which was unknown to us at the time the main results of the present paper were obtained) some necessary and sufficient conditions for solving the DDPU have been established. More precisely, the result of [17] says (in our terminology) that the DDPU is solvable for a quintuple $(E, F, G, D, H) \in \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{X}, \mathcal{Z}) \times \mathcal{L}(\mathcal{U}, \mathcal{Z}) \times \mathcal{L}(\mathcal{D}, \mathcal{Z}) \times \mathcal{L}(\mathcal{X}, \mathcal{Y})$ if and only if the system $\tilde{\mathfrak{S}}_d(E, F, G)$ is U-regularizable and there exist subspaces $\mathcal{V}$, $\mathcal{W} \subset \operatorname{Ker} H$ and $\mathcal{P} \subset \mathcal{Z}$ such that the following conditions hold:

(i) $\mathcal{V}$ is $\tilde{\mathfrak{S}}(E, F, G)$-invariant;
(ii) $\mathcal{V} \cap \operatorname{Ker} E = 0$;
(iii) $\mathcal{W}$ is $\tilde{\mathfrak{S}}(E, F, G)$-invariant;
(iv) $\dim(E\mathcal{W} \cap \operatorname{Im} G) \leqq \dim((F^{-1} \operatorname{Im} G) \cap \mathcal{W})$;
(v) $\operatorname{Im} D \subset E\mathcal{V} + \mathcal{P}$;
(vi) $E\mathcal{W} \subset \mathcal{P} \subset F\mathcal{W} + \operatorname{Im} G$;
(vii) $\dim(\mathcal{P} \cap \operatorname{Im} G) \leqq \dim((F^{-1} \operatorname{Im} G) \cap \mathcal{W})$.

Note that the problem of finding spaces $\mathcal{V}$, $\mathcal{W}$, and $\mathcal{P}$ satisfying the above conditions has not been solved in [17] and seems to be difficult. It should be contrasted with the explicit conditions of Theorem 5.3. More importantly, [17] does not provide any hint of how to construct a feedback map solving the DDPU.

Remark 5.3. Together with a given quintuple (E, F, G, H, D) consider also $(\tilde{E}, \tilde{F}, \tilde{G}, \tilde{H}, \tilde{D}) := (PEQ, P(F+GM)Q, PG, HQ, PD)$, where P, Q are invertible and M is an arbitrary feedback map. Then it is obvious that the DDP (or DDPU) is solvable for the quintuple (E, F, G, H, D) if and only if the problem is solvable for $(\tilde{E}, \tilde{F}, \tilde{G}, \tilde{H}, \tilde{D})$. In other words, the property of solvability of the DDP and DDPU is invariant with respect to the action of the above-described group of transformations. In particular, the property is invariant under the (Rosenbrock) restricted system equivalence (cf. [32]). However, if the (Verghese) strong system equivalence (cf. [32]) is taken into account, the situation is not apparent since strong system equivalence does not preserve the class of systems considered in our paper. As is easily seen, the orbit of an implicit system (under the strong system equivalence) may contain a system with a direct influence of the input u_k and the disturbance z_k on the output y_k. Unfortunately, the methods of our paper are not directly applicable to solving the DDP (or DDPU) for such systems. The problem of disturbance decoupling for systems with direct dependence of the output on the input and disturbance remains open.

TABLE 1

The present paper and [2]-[9]	Özcaldiran, Lewis and Malabre [25], [26], [21]	Willems [33]
$\vec{\mathcal{V}}(E, F, G)\|\mathcal{W}$	$\mathbf{V}^*(E, F, G; \mathcal{W})$	$\mathbf{V}^*_\omega$
greatest $\vec{\mathfrak{S}}(E, F, G)$-invariant subspace contained in $\mathcal{W}$	supremal (F, E, G)-invariant subspace of $\mathcal{W}$	supremal controlled invariant subspace contained in $\mathcal{W}$
$\overleftarrow{\mathcal{V}}(E, F, G)\|\mathcal{W}$	$\mathbf{V}^*(F, E, G; \mathcal{W})$	—
greatest $\overleftarrow{\mathfrak{S}}(E, F, G)$-invariant subspace contained in $\mathcal{W}$	supremal (E, F, G)-invariant subspace of $\mathcal{W}$	
$\vec{\mathcal{R}}(E, F, G)\|\mathcal{W}$	$\mathbf{R}^*_a(E, F, G; \mathcal{W})$	$\mathbf{R}^*_{a,\omega}$
greatest $\vec{\mathfrak{S}}(E, F, G)$-reachability subspace contained in $\mathcal{W}$	supremal almost reachability subspace of $\mathcal{W}$	supremal almost controllability subspace contained in $\mathcal{W}$
$\vec{\mathcal{C}}(E, F, G)\|\mathcal{W}$	$\mathbf{R}^*(E, F, G; \mathcal{W})$	$\mathbf{R}^*_\omega$
greatest $\vec{\mathfrak{S}}(E, F, G)$-controllability subspace contained in $\mathcal{W}$	supremal reachability subspace of $\mathcal{W}$	supremal controllability subspace contained in $\mathcal{W}$

We can observe that when $E = I$ the DDP, DDPU, and the disturbance decoupling problem considered in [36, Chap. 4] are equivalent. To show this it is sufficient (in view of [36, Chap. 4]) to prove the following simple result.

PROPOSITION 5.2. *Assume that* $\mathcal{Z} = \mathcal{X}$ *and* $(A, B, D, C) \in \mathcal{L}(\mathcal{X}, \mathcal{X}) \times \mathcal{L}(\mathcal{U}, \mathcal{X}) \times \mathcal{L}(\mathcal{D}, \mathcal{X}) \times \mathcal{L}(\mathcal{X}, \mathcal{Y})$. *Then the* DDP *is solvable for the quintuple* (I, A, B, D, C) *if and only if* $\operatorname{Im} D \subset \vec{\mathcal{V}}(I, A, B)|\operatorname{Ker} H$.

Proof. ($\Rightarrow$) Let $K \in \mathcal{L}(\mathcal{X}, \mathcal{U})$ be such that the system $\vec{\mathfrak{S}}_d(I, A+BK, 0)$ accepts all disturbance sequences from $\sigma_0^+(\operatorname{Im} D)$. Then by Proposition 1.5 $\operatorname{Im} D \subset \vec{\mathcal{V}}(I, A+BK, 0)|\operatorname{Ker} H + A\vec{\mathcal{R}}(I, A+BK, 0)|\operatorname{Ker} H$. But $\vec{\mathcal{R}}(I, A+BK, 0) = 0$.

($\Leftarrow$) The proof is an obvious consequence of Theorem 4.2 of [36]. □

Appendix. We can show (cf. [3], [4], [7]) that the spaces $\vec{\mathcal{V}}(E, F, G)|\mathcal{W}$, $\tilde{\mathcal{V}}(E, F, G)|\mathcal{W}$, $\vec{\mathcal{R}}(E, F, G)|\mathcal{W}$, and $\vec{\mathcal{C}}(E, F, G)|\mathcal{W}$ can be computed with the aid of some (finite) subspace recursions. The recursions coincide with those considered in [23]-[26] (cf. also [21]). Consequently, we can identify these spaces with some spaces considered in the above-mentioned references. However the terminology and notation used in [21], [23]-[26] differs from that used in the present paper (and also [2]-[9]). Table 1 will make the access to our results easier for some readers. It summarizes the basic terminology of various references.

The above-reported terminology of [21], [25], [26], and [33] concerns continuous-time systems. Moreover, in [33] only standard systems (i.e., with $E = I$) are considered. At this point it is worth noting that Willems in [34, § 8] (in the context of discrete-time standard systems) proposes for counterparts of spaces $\mathbf{V}_{\omega}^{*}$, $\mathbf{R}_{a,\omega}^{*}$, and $\mathbf{R}_{\omega}^{*}$ other symbols and names.

Acknowledgments. We are very thankful to the reviewers of this paper for various corrections and helpful suggestions. We are also grateful to Professor L. R. Fletcher for making his paper available to us.

REFERENCES

[1] J. D. APLEVICH, *Minimal representations of implicit linear systems*, Automatica, 21 (1985), pp. 259-269.
[2] A. BANASZUK, *Analysis of implicit linear discrete-time systems*, Ph.D. thesis, Department of Electrical Engineering, Warsaw University of Technology, Warsaw, Poland, 1989. (In Polish.)
[3] A. BANASZUK, M. KOCIĘCKI, AND K. M. PRZYŁUSKI, *Implicit linear discrete-time systems*, Mathematics of Control, Signals and Systems, 3 (1990), to appear, see also Preprint 397 (1987), Institute of Mathematics, Polish Academy of Science, Warsaw, Poland.
[4] ———, *Remarks on controllability of implicit linear discrete-time systems*, Systems Control Lett., 10 (1988), pp. 67-70.
[5] ———, *On almost invariant subspaces for implicit linear discrete-time systems*, Systems Control Lett., 11 (1988), pp. 289-297.
[6] ———, *On duality between observation and control for implicit linear discrete time systems*, submitted.
[7] ———, *Remarks on the theory of implicit linear discrete-time systems*, in Proc. 1987 Internat. Symposium on Singular Systems, Atlanta, GA, December 1987, F. L. Lewis, ed., Georgia Tech. Research Corp., Atlanta, GA. pp. 44-47.
[8] ———, *On Hautus-type conditions for controllability of implicit linear discrete-time systems*, Circuits Systems Signal Process., 8 (1989), pp. 289-298.
[9] ———, *On Kalman-type decomposition for implicit linear discrete-time systems and its applications*, Internat. J. Control, to appear.
[10] D. J. BENDER, *The disturbance decoupling problem for descriptor systems*, unpublished manuscript.
[11] P. BERNHARD, *On singular implicit linear dynamical systems*, SIAM J. Control Optim., 20 (1982), pp. 612-633.
[12] J. E. BOYARINCEV, *Regular and Singular Systems of Linear Ordinary Differential Equations*, Nauka, Novosibirsk, 1980. (In Russian.)
[13] S. L. CAMPBELL, *Singular Systems of Differential Equations*, Pitman, San Francisco, Vol. 1 1980; Vol. 2, 1982.
[14] P. M. COHN, *Algebra*, Vol. 2, John Wiley, London, 1977.
[15] B. DZIURLA AND R. W. NEWCOMB, *Nonregular semistate systems: examples and input-output pairing*, in Proc. 26th Annual IEEE Conference on Decision and Control, IEEE Computer Society, New York, 1987, pp. 1125-1126.
[16] L. R. FLETCHER, *Regularizability of descriptor systems*, Internat. J. Systems Sci., 17 (1986), pp. 843-847.
[17] L. R. FLETCHER AND A. AASARAAI, *On disturbance decoupling in descriptor systems*, SIAM J. Control Optim., 27 (1989), pp. 1319-1332.
[18] F. L. LEWIS, *A survey of linear singular systems*, Circuits Systems Signal Process., 5 (1986), pp 3-36.

[19] D. G. LUENBERGER, *Dynamical systems in descriptor form*, IEEE Trans. Automat. Control, 22 (1977), pp. 312-321.

[20] D. G. LUENBERGER AND A. ARBEL, *Singular dynamic Leontief model*, Econometrica, 45 (1977), pp. 991-995.

[21] M. MALABRE, *More geometry about singular systems*, in Proc. 26th Annual IEEE Conference on Decision and Control, IEEE Computer Society, New York, 1987, pp. 1138-1139.

[22] R. NIKOUKHAH, A. S. WILLSKY, AND B. C. LEVY, *Boundary-value descriptor systems: well-posedness, reachability and observability*, Internat. J. Control, 46 (1987), pp. 1715-1737.

[23] K. ÖZCALDIRAN, *Control of descriptor systems*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 1985.

[24] ———, *A geometric characterization of the reachable and the controllable subspace of descriptor systems*, Circuits Systems Signal Process., 5 (1986), pp. 37-48.

[25] ———, *Geometric notes on descriptor systems*, in Proc. 26th Annual IEEE Conference on Decision and Control, IEEE Computer Society, New York, 1987, pp. 1134-1137.

[26] K. ÖZCALDIRAN AND F. L. LEWIS, *On the regularizability of singular systems*, IEEE Trans. Automat. Control, 35 (1990), to appear.

[27] M. A. SHAYMAN AND Z. ZHOU, *Feedback control and classification of generalized linear systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 483-494.

[28] G. W. STEWART, *On the sensitivity of the eigenvalue problem* $Ax = \lambda Bx$, SIAM J. Numer. Anal., 9 (1972), pp. 669-686.

[29] H. L. TRENTELMAN, *Almost Invariant Subspaces and High Gain Feedback*, CWI, Amsterdam, 1986.

[30] P. VAN DOOREN, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, 26 (1981), pp. 111-129.

[31] ———, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121-135.

[32] G. C. VERGHESE, B. LEVY, AND T. KAILATH, *A generalized state space for singular systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 811-831.

[33] J. C. WILLEMS, *Almost invariant subspaces: an approach to high gain feedback design—part* I: *almost controlled invariant subspaces*, IEEE Trans. Automat. Control, 26 (1981), pp. 235-252.

[34] ———, *Almost invariant subspaces: an approach to high gain feedback design—part* II: *almost conditionally invariant subspaces*, IEEE Trans. Automat. Control, 26 (1982), pp. 1071-1084.

[35] K. T. WONG, *The eigenvalue problem* $\lambda Tx + Sx$, J. Differential Equations, 16 (1974), pp. 270-280.

[36] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1294–1320, November 1990

003

DIFFERENTIAL GAMES: A VIABILITY APPROACH*

JEAN-PIERRE AUBIN†

Abstract. The usual intertemporal optimality criterion traditionally used in differential games is replaced here by a myopic criterion, playability, which requires that at each instant, the state of the game obeys playability constraints. (For simplicity, only time-independent playability constraints are presented below.) Game theoretical concepts are adapted to this case and characterized through conveniently generalized Isaacs' equations, the contingent Isaacs' inequalities.

For each of these concepts, feedback controls are constructed, according to several game theoretical selection procedures when they are not uniquely determined by contingent Isaacs' inequalities.

The question of choosing strategies through their velocities regarded as decisions is also investigated, and decision rules allowing victory or defeat are characterized through other contingent partial differential equations.

Key words. differential games, viability theory, playable games, viability kernels, heavy trajectories, minimal

AMS(MOS) subject classifications. 90025, 90D26

Introduction. We consider a two-player differential game whose dynamics are described by

(a) (i) $x'(t) = f(x(t), y(t), u(t))$,
(ii) $u(t) \in U(x(t), y(t))$,
(b) (i) $y'(t) = g(x(t), y(t), v(t))$,
(ii) $v(t) \in V(x(t), y(t))$,

where u, v, the controls, are regarded as *strategies* used by the players to govern the evolution of the states x, y of the game.

The *rules of the game* are set-valued maps $P: Y \rightsquigarrow X$ and $Q: X \rightsquigarrow Y$, describing the constraints imposed by one player on the other. They replace the traditional intertemporal optimality or endpoint criteria used in differential games.

The *playability domain* of the game $K \subset X \times Y$ is defined by

$$K := \{(x, y) \in X \times Y \mid x \in P(y) \text{ and } y \in Q(x)\}.$$

(We consider only the time-independent case for the sake of simplicity. See [42] for the extension to time-dependent problems.) We point out the following properties:

THE PLAYABILITY PROPERTY. It states that for any initial state $(x_0, y_0) \in K$, there exists a solution to the differential game which is playable in the sense that

$$\forall t \geqq 0, \quad x(t) \in P(y(t)) \quad \text{and} \quad y(t) \in Q(x(t)).$$

XAVIER'S DISCRIMINATING PROPERTY. It states that for any initial state $(x_0, y_0) \in K$ and for any continuous closed-loop strategy $\tilde{v}(\cdot, \cdot)$ played by Yvette, there exists a playable solution to the differential game.

XAVIER'S LEADING PROPERTY. It states that there exists a continuous closed-loop strategy $\tilde{u}(\cdot, \cdot)$ played by Xavier such that for any initial state $(x_0, y_0) \in K$, there exists a playable solution to the differential game.

Our first task is to characterize the rules satisfying such properties as somewhat generalized solutions to Isaacs' equations. Since the rules are set-valued maps and not

* Received by the editors September 12, 1988; accepted for publication (in revised form) October 18, 1989.
† Centre de Recherches de Mathématiques de la Décision, Université de Paris-Dauphine, Paris, France.

functions, we characterize them by the indicators Ψ_P and Ψ_Q of their graphs, defined by $\Psi_P(x, y) := 0$ when $x \in P(y)$ and $\Psi_P(x, y) := +\infty = 0$ when $x \notin P(y)$. But these functions, which are only lower semicontinuous (when the graphs are closed) are not differentiable in the usual sense. Hence we must replace the concept of derivative by the one of contingent epiderivative[1] in the Isaacs equations.

This being done, we shall interpret the solutions to contingent Isaacs' equations in game theoretical terms and characterize the above properties of the rules P and Q by checking whether the function $\max(\Psi_P, \Psi_Q)$ is a solution to the corresponding contingent Isaacs' equation.

We focus our attention in the second section to the playability property.

We shall characterize it by constructing *retroaction rules*

$$(x, y, v) \rightsquigarrow C(x, y; v) \quad \text{and} \quad (x, y, u) \rightsquigarrow D(x, y; u),$$

which involve the contingent derivatives of the set-valued maps P and Q, with which we build the *regulation map* R mapping each $(x, y) \in K$ to the regulation set

$$R(x, y) = \{(u, v) \mid u \in C(x, y; v) \text{ and } v \in D(x, y; u)\}.$$

The strategies belonging to $R(x, y)$ are called *playable*.

The Playability Theorem states that under technical assumptions, the playability property holds true if and only if

$$\forall (x, y) \in K, \quad R(x, y) \neq \varnothing$$

and that playable solutions to the game are regulated by the *regulation law*:

$$\forall t \geqq 0, \quad u(t) \in C(x(t), y(t); v(t)) \quad \text{and} \quad v(t) \in D(x(t), y(t); u(t)).$$

We then deal in § 3 with the construction of single-valued *playable feedbacks* $(\tilde{u}, \tilde{v})$, such that the differential system

$$x'(t) = f(x(t), y(t), \tilde{u}(x(t), y(t)),$$
$$y'(t) = g(x(t), y(t), \tilde{v}(x(t), y(t)),$$

has playable solutions for each initial state. By the Playability Theorem, they must be selections of the regulation map R in the sense that

$$\forall (x, y) \in K, \quad (x, y) \mapsto (\tilde{u}(x, y), \tilde{v}(x, y)) \in R(x, y).$$

---

[1] We recall that the *contingent cone* $T_K(x)$ to a subset K at $x \in K$ is the closed cone of elements v satisfying

$$\liminf_{h \to 0+} d(x + hv, K)/h = 0.$$

The contingent epiderivative $D_\uparrow V(x)$ of an extended function V from X to $\mathbf{R} \cup \{+\infty\}$ at $x \in \mathrm{Dom}(V)$ is defined by

$$\mathscr{E}pD_\uparrow V(x) := T_{\mathscr{E}p(V)}(x, V(x))$$

or, equivalently, by

$$D_\uparrow V(x)(u) = \liminf_{h \to 0+, u' \to u} \frac{V(x + hu') - V(x)}{h}.$$

The *contingent derivative* of the set-valued map Q from X to Y at a point (x, y) of its graph is the closed positively homogenous set-valued map $DQ(x, y)$ from X to Y defined by

$$\mathrm{Graph}(DQ(x, y)) := T_{\mathrm{Graph}(Q)}(x, y)$$

or, equivalently, by

$$v \in DQ(x, y)(u) \Leftrightarrow \liminf_{h \to 0+, u' \to u} d\left(v, \frac{Q(x + hu') - y}{h}\right) = 0.$$

We shall prove the existence of such continuous single-valued playable feedbacks, as well as more constructive, but discontinuous, playable feedbacks, such as the feedbacks associating the strategies of $R(x, y)$ with minimal norm (the playable slow feedbacks, as in [21], [47]). More generally, we shall show the existence of possibly set-valued feedbacks associating with any $(x, y) \in K$ the set of strategies $(u, v) \in R(x, y)$ which are solutions to a (static) optimization problem of the form:

$$(u, v) \in R(x, y) \,|\, \sigma(x, y; u, v) \leqq \inf_{u', v' \in R(x,y)} \sigma(x, y; u', v')$$

or solutions to a noncooperative game of the form:

$$\forall (u', v') \in R(x, y), \quad a(x, u, v') \leqq a(x, u, v) \leqq a(x, u', v).$$

In other words,

> the players can implement playable solutions to the differential game by playing for each state $(x, y) \in K$ a static game on the strategies of the regulation subset $R(x, y)$.

We also consider in § 4 the issue of finding *discriminating feedbacks* by providing, for instance, sufficient conditions implying that for all continuous feedback $\tilde{v}(x, y) \in V(x, y)$ played by Yvette, Xavier can find a feedback (continuous or of minimal norm) $\tilde{u}(x, y)$ such that the differential equation above has playable solutions for each initial state.

We address the question of whether Xavier has a leading role, i.e., the problem of constructing continuous *pure feedbacks* $\tilde{u}(x, y)$ which have the property of yielding playable solutions to the above differential game whatever the strategy played by Yvette.

The last section is devoted to closed-loop decision rules, which *operate on the velocities of the strategies* (regarded as *decisions*) rather than on the controls. We need to provide first regulation maps which yield absolutely continuous strategies which are then almost everywhere differentiable. We distinguish among them those that guarantee or allow victory or defeat adequately defined. The indicator functions of their graphs are characterized as solutions of contingent partial differential inequalities. We apply analogous selection procedures which yield closed-loop decision rules allowing, say, a game to remain stable.

The techniques rely heavily on viability theorems and differential calculus of set-valued maps, which are exposed in [2, Chaps. 4, 5, and 6] and [6]. An Appendix presents some results on lower semicontinuous Lyapunov functions which are used in some proofs. The time-dependent case and some classical examples appear in [42].

1. Contingent version of Isaacs' equations. Let us consider only two players, Xavier and Yvette. Xavier acts on a state space X and Yvette on a state space Y. For doing so, they have access to some knowledge about the global state (x, y) of the system and are allowed to choose strategies u in a global state-dependent set $U(x, y)$ and v in a global state-dependent set $V(x, y)$, respectively.

Their actions on the state of the system are governed by the system of differential inclusions:

$$\begin{array}{lll} (1) & \text{(a)} & \text{(i)} \quad x'(t) = f(x(t), y(t), u(t)), \\ & & \text{(ii)} \quad u(t) \in U(x(t), y(t)), \\ & \text{(b)} & \text{(i)} \quad y'(t) = g(x(t), y(t), v(t)), \\ & & \text{(ii)} \quad v(t) \in V(x(t), y(t)). \end{array}$$

We now describe the influences (power relations) that Xavier exerts on Yvette and vice versa through *rules of the game.* They are set-valued maps $P: Y \rightsquigarrow X$ and $Q: X \rightsquigarrow Y$ which are interpreted in the following way. When the state of Yvette is y, Xavier's choice is constrained to belong to $P(y)$. In a symmetric way, the set-valued map Q assigns to each state x the set $Q(x)$ of states y that Yvette can implement.

Hence, the *playability subset* of the game is the subset $K \subset X \times Y$ defined by

$$K := \{(x, y) \in X \times Y \mid x \in P(y) \text{ and } y \in Q(x)\}. \tag{2}$$

Naturally, we must begin by providing sufficient conditions implying that the playability subset is nonempty. Since the playability subset is the subset of fixed points (x, y) of the set-valued map $(x, y) \rightsquigarrow P(y) \times Q(x)$, we can use one of the many fixed point theorems to answer these types of questions.[2]

From now on, *we shall assume that the playability subset associated with the rules P and Q is not empty.*

We can reformulate this differential game in a more compact form, by denoting by $z := (x, y) \in Z := X \times Y$ the global state, by $h(z, u, v) := (f(x, u, v), g(y, u, v))$ the values of the map $h: \mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q \to \mathbf{R}^n$ describing the dynamics of the game, by $L := \text{Graph}(P)$ Xavier's closed domain of definition, by $M := \text{Graph}(Q^{-1})$ Yvette's, and by $K := L \cap M$ the playability subset. We shall also identify the set-valued maps U and V with their restrictions to L and M, respectively, by setting $U(z) := \varnothing$ whenever $z \notin L$ and $V(z) := \varnothing$ when $z \notin M$.

Hence the differential game can be written in the form

$$\begin{aligned} &\text{(i)} \quad z'(t) = h(z(t), u(t), v(t)), \\ &\text{(ii)} \quad u(t) \in U(z(t)), \\ &\text{(iii)} \quad v(t) \in V(z(t)). \end{aligned} \tag{3}$$

We denote by $\mathcal{S}(z_0)$ the subset of solutions $z(\cdot)$ to (3) starting at z_0.

Let us associate with this differential game the following four Hamilton-Jacobi-Isaacs partial differential equations:

(i) $\inf_{u \in U(z)} \inf_{v \in V(z)} d\Phi(z)/dz \cdot h(z, u, v) = 0$,
(ii) $\sup_{u \in U(z)} \sup_{v \in V(z)} d\Phi(z)/dz \cdot h(z, u, v) = 0$,
(iii) $\sup_{v \in V(z)} \inf_{u \in U(z)} d\Phi(z)/dz \cdot h(z, u, v) = 0$,
(iv) $\inf_{u \in U(z)} \sup_{v \in V(z)} d\Phi(z)/dz \cdot h(z, u, v) = 0$.

We would like to study the properties of the solutions to these partial differential equations, and in particular, characterize the solutions which are indicators of closed subsets L, defined by

$$\Psi_L(x) := \begin{cases} 0 & \text{if } z \in L, \\ +\infty & \text{if } z \notin L \end{cases}$$

and which are only lower semicontinuous.

Hence we are led to weaken the concept of usual derivatives involved in these partial differential equations by replacing them by contingent epiderivatives, since any extended function $\Phi: X \to \mathbf{R} \cup \{+\infty\}$ has contingent epiderivative, and in particular, indicators, for which we have the relation

$$D_\uparrow \Psi_L(z)(v) = \Psi_{T_L(z)}(v) := \begin{cases} 0 & \text{if } v \in T_L(z), \\ +\infty & \text{if } v \notin T_L(z). \end{cases}$$

[2] For instance, Kakutani's fixed point theorem (see [6, Chap. 3]) furnishes such conditions: let $L \subset X$ and $M \subset Y$ be compact convex subsets and $P: M \rightsquigarrow L$ and $Q: L \rightsquigarrow M$ be closed maps with nonempty convex images. Then the playability subset is not empty.

THEOREM 1.1. *Let us assume at least that* $h : \mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q \to \mathbf{R}^n$ *is continuous, has linear growth, and that the set-valued maps are closed with linear growth.*

We assume that all extended functions Φ *are nonnegative and contingently epidifferentiable*[3] *and that their domains are contained in the intersection* K *of the domains of* U *and* V.

(1) *If the values of the set-valued maps* U *and* V *are convex and if* h *is affine with respect to the controls,* Φ *is a solution to the contingent equation*

$$\inf_{u \in U(z)} \inf_{v \in V(z)} D_\uparrow \Phi(z)(h(z, u, v)) \leqq 0 \tag{4}$$

if and only if

$$\forall z_0 \in \mathrm{Dom}\,(\Phi), \quad \exists z(\cdot) \in \mathscr{S}(z_0) \,|\, \forall t \geqq 0, \quad \Phi(z(t)) \leqq \Phi(z).$$

(2) *Assume that* h *is uniformly Lipschitzean with respect to* x. *Then* Φ *is a solution to the contingent equation*

$$\sup_{u \in U(z)} \sup_{v \in V(z)} D_\uparrow \Phi(z)(h(z, u, v)) \leqq 0 \tag{5}$$

if and only if

$$\forall z_0 \in \mathrm{Dom}\,(\Phi), \quad \forall z(\cdot) \in \mathscr{S}(z_0), \quad \forall t \geqq 0, \quad \Phi(z(t)) \leqq \Phi(z_0).$$

(3) *Assume that* V *is lower semicontinuous, that the values of* U *and* V *are convex, and that* h *is affine with respect to* u. *Then* Φ *is a solution to the contingent equation*

$$\sup_{v \in V(z)} \inf_{u \in U(z)} D_\uparrow \Phi(z)(h(z, u, v)) \leqq 0 \tag{6}$$

if and only if for any continuous closed-loop strategy $\tilde{v}(z) \in V(z)$ *played by Yvette and any initial state* $z_0 \in \mathrm{Dom}\,(\Phi)$, *there exists a solution* $z(\cdot)$ *to Xavier's control problem*

(i) $z'(t) = hz(t), u(t), \tilde{v}(z(t)))$,

(ii) $u(t) \in U(z(t))$

starting at z_0 *and satisfying for all* $t \geqq 0$, $\Phi(z(t)) \leqq \Phi(z_0)$.

(4) *Assume that* V *is lower semicontinuous with convex values. Then* Φ *is a solution to the contingent equation*

$$\inf_{u \in U(z)} \sup_{v \in V(z)} D_\uparrow \Phi(z)(h(z, u, v)) \leqq 0 \tag{7}$$

if and only if Xavier can play a closed-loop strategy $\tilde{u}(z) \in U(z)$ *such that, for any continuous closed-loop strategy* $\tilde{v}(z) \in V(z)$ *played by Yvette and for any initial state* $z_0 \in \mathrm{Dom}\,(\Phi)$, *there exists a solution* $z(\cdot)$ *to*

$$z'(t) = h(z(t), \tilde{u}(z(t)), \tilde{v}(z(t)) \tag{8}$$

starting at z_0 *and satisfying for all* $t \geqq 0$, $\Phi(z(t)) \leqq \Phi(z_0)$. *The converse is true if*

$$B_\Phi(z) := \{\bar{u} \in U(z) \text{ such that } \sup_{v \in V(z)} D_\uparrow \Phi(z)(h(z, \bar{u}, v)) = \inf_{u \in U(z)} \sup_{v \in V(z)} D_\uparrow \Phi(z)(h(z, u, v))\}$$

is lower semicontinuous with closed convex values.

Proof.

— The two first statements are translations of the theorems characterizing Lyapunov and universal Lyapunov functions (see the Appendix) applied to the differential inclusion $z'(t) \in H(z(t))$ where $H(z) := f(z, U(z), V(z))$.

[3] This means that for all $z \in \mathrm{Dom}\,(\Phi)$, for all $v \in X$, $D_\uparrow \Phi(z)(v) > -\infty$ and that $D_\uparrow \Phi(z)(v) < \infty$ for at least a $v \in X$.

— Let us prove the third one. Assume that Φ satisfies the stated property. Since V is lower semicontinuous with convex values, Michael's theorem (see [2, Chap. 1]) implies that for all $z_0 \in \mathrm{Dom}(V)$ and $v_0 \in V(z_0)$, there exists a continuous selection $\tilde{v}(\cdot)$ of V such that $v(z_0) = v_0$. Then Φ enjoys the Lyapunov property for the set-valued map $H_{\tilde{v}}(z) := h(z, U(z), \tilde{v}(z))$, and thus, there exists $u_0 \in U(z_0)$ such that

$$D_{\uparrow}\Phi(z_0)(h(z_0, u_0, \tilde{v}(z_0))) \leqq 0.$$

Hence Φ is a solution to (6).

Conversely, assume that Φ is a solution to (6). Then for any closed-loop strategy $\tilde{v}$, the set-valued map $H_{\tilde{v}}$ satisfies the assumptions of the theorem characterizing Lyapunov functions, so that there exists a solution to the inclusion $z' \in H_{\tilde{v}}(z)$ for any initial state $z \in \mathrm{Dom}(\Phi)$ satisfying for all $t \geqq 0$, $\Phi(z(t)) \leqq \Phi(z)$.

— Consider finally the fourth statement. Assume that Xavier can find a continuous closed-loop strategy $\tilde{u}$ such that for any closed-loop strategy $\tilde{v}$, Φ enjoys the stated property. Since V is lower semicontinuous with convex values, Michael's theorem implies that for all $z_0 \in \mathrm{Dom}(V)$ and $v_0 \in V(z_0)$, there exists a continuous selection $\tilde{v}(\cdot)$ of V such that $v(z_0) = v_0$. Since for any continuous closed-loop strategy $\tilde{v}(\cdot)$, Φ enjoys the Lyapunov property for the single-valued map $z \to h(z, \tilde{u}(z), \tilde{v}(z))$, we deduce that for all $z_0 \in \mathrm{Dom}(\Phi)$, there exists $u := \tilde{u}(z)$ such that for all $v \in V(z)$, $D_{\uparrow}\Phi(z)(h(x, u, v)) \leqq 0$, so that Φ is a solution to (6).

Conversely, assume that the set-valued map B_{Φ} is lower semicontinuous with closed convex values. Hence Michael's theorem implies that there exists a continuous selection $\tilde{u}$ of B_{Φ}. Then for any continuous closed-loop strategy $\tilde{v}(\cdot) \in V(\cdot)$, we deduce from (7) that Φ is a Lyapunov function for the single-valued map $z \to h(z, \tilde{u}(z), \tilde{v}(z))$, so that, for all $z \in \mathrm{Dom}(\Phi)$, there exists a solution $z(\cdot)$ to the system (8) satisfying for all $t \geqq 0$, $\Phi(z(t)) \leqq \Phi(z)$. □

Let L be a closed subset of the intersection K of the domains of U and V. The problem we investigate is that of finding one (or all) solution(s) $z(\cdot)$ of the game which is (are) viable in L. There are several ways to achieve that purpose, according to the cooperative or noncooperative behavior of the players. Here we shall investigate several of them.

DEFINITION 1.1. We shall say the subset L enjoys:

(1) The "playability property" if and only if

$$\forall z \in L, \quad \exists z(\cdot) \in \mathscr{S}(z) \,|\, \forall t \geqq 0, \quad z(t) \in L.$$

(2) The "winability property" if and only if

$$\forall z \in L, \quad \forall z(\cdot) \in \mathscr{S}(z), \quad \forall t \geqq 0, \quad z(t) \in L.$$

(3) "Xavier's discriminating property" if and only if for any continuous closed-loop strategy $\tilde{v}(z) \in V(z)$ played by Yvette and any initial state $z \in L$, there exists a solution $z(\cdot)$ to Xavier's control problem

(i) $z'(t) = h(z(t), u(t), \tilde{v}(z(t)))$,

(ii) $u(t) \in U(z(t))$

starting at z and which is viable in L.

(4) "Xavier's leading property" if and only if Xavier can play a closed-loop strategy $\tilde{u}(z) \in U(z)$ such that, for any continuous closed-loop strategy $\tilde{v}(z) \in V(z)$ played by Yvette and for any initial state $z \in L$, there exists a solution $z(\cdot)$ to (8) starting at z and viable in L.

We shall characterize these properties: for that purpose we associate with L the following set-valued maps:

— The regulation map R_L defined by

$$\forall z \in L, \quad R_L(z) := \{(u, v) \in U(z) \times V(z) \mid h(z, u, v) \in T_L(z)\};$$

— Xavier's discriminating map A_L defined by

$$\forall z \in L, \quad A_L(z, v) := \{u \in U(z) \mid (u, v) \in R_L(z)\};$$

— Xavier's leading map B_L defined by

$$\forall z \in L, \quad B(z) := \bigcap_{v \in V(z)} A_L(z, v).$$

DEFINITION 1.2. We shall say that

— L is a playability domain if for all $z \in L$, $R_L(z) \neq \emptyset$.

— L is a winability domain if for all $z \in L$, $R_L(z) := U(z) \times V(z)$.

— L is Xavier's discriminating domain if

$$\forall z \in L, \quad \forall v \in V(z), \quad A_L(z, v) \neq \emptyset. \tag{9}$$

— L is Xavier's leading domain if for all $z \in L$, $B_L(z) \neq \emptyset$.

We begin by translating these properties in terms of contingent version of Isaacs' equations.

PROPOSITION 1.1. *Let us assume that $h : \mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q \to \mathbf{R}^n$ is continuous, has linear growth, and that the set-valued maps are closed with linear growth.*

— *L is playability domain if and only if Ψ_L is a solution to* (4).

— *L is a winability domain if and only if Ψ_L is a solution to* (5).

— *L is a discriminating domain for Xavier if and only if Ψ_L is a solution to* (6).

— *L is a leading domain for Xavier if and only if Ψ_L is a solution to* (7).

Therefore, Theorem 1.1 implies the following characterization of these domains.

COROLLARY 1.1. *Let us assume at least that $h : \mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q \to \mathbf{R}^n$ is continuous, has linear growth, and that the set-valued maps are closed with linear growth.*

(1) *If the values of the set-valued maps U and V are convex and if h is affine with respect to the controls, then L enjoys the playability property if and only if it is a playability domain.*

(2) *Assume that h is uniformly Lipschitzean with respect to x. Then L enjoys the winability property if and only if it is a winability domain.*

(3) *Assume that V is lower semicontinuous, that the values of U and V are convex and that h is affine with respect to u. Then L enjoys Xavier's discriminating property if and only if it is a discriminating domain for Xavier.*

(4) *Assume that V is lower semicontinuous with convex values. If L enjoys Xavier's leading property, then it is a leading domain for him. The converse is true if B_L is lower semicontinuous with closed convex values.*

The existence theorems of the viability and invariance kernels imply the following consequence.

PROPOSITION 1.2. *Let us assume that $h : \mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q \to \mathbf{R}^n$ is continuous, has linear growth, and that the set-valued maps are closed with linear growth.*

(1) *If the values of the set-valued maps U and V are convex and if h is affine with respect to the controls, then there exists a largest closed playability domain contained in L, whose indicator is the smallest lower semicontinuous solution to* (4) *larger than or equal to the indicator Ψ_L of L.*

(2) *Assume that h is uniformly Lipschitzean with respect to x. Then there exists a largest closed winability domain contained in L, whose indicator is the smallest lower semicontinuous solution to* (5) *larger than or equal to the indicator* Ψ_L *of L.*

2. Playable differential games. We shall now proceed with the case of the game described by (1), where the playability domain is defined from rules P and Q by

$$K := \{(x, y) \in X \times Y \mid x \in P(y) \text{ and } y \in Q(x)\}$$

enjoys the *playability property*, which becomes in this case: for any initial state $(x_0, y_0) \in K$, there exists a solution to the differential game (1) which is *playable* in the sense that

$$\forall t \geqq 0, \quad x(t) \in P(y(t)) \quad \text{and} \quad y(t) \in Q(x(t)).$$

We now need to define *playable rules.* For that purpose, we associate with the rules P and Q acting on the states *retroaction rules* C and D acting on the strategies defined in the following way.

DEFINITION 2.1. Xavier's *retroaction rule* is the set-valued map C defined by

$$C(x, y; v) = \{u \in U(x, y) \mid f(x, y, u) \in DP(y, x)(g(x, y, v))\}$$

and Yvette's *retroaction rule* is the set-valued map D defined by

$$D(x, y; u) = \{v \in V(x, y) \mid g(x, y, v) \in DQ(x, y)(f(x, y, u))\}.$$

We associate with them the *regulation map* R defined by

$$R(x, y) = \{(u, v) \mid u \in C(x, y; v) \text{ and } v \in D(x, y; u)\}. \tag{10}$$

The subset $R(x, y)$ is called the *regulation set* and its elements *playable controls.*

In other words, we have associated with each state (x, y) of the playability domain a static game on the strategies defined by the retroaction rules. This new game on strategies is playable if the subset $R(x, y)$ is nonempty. This property deserves a definition.

DEFINITION 2.2. We shall say that P and Q are *playable rules* if their graphs are closed, the playability domain K defined by (2) is nonempty, and if for all pairs $(x, y) \in K$, the values $R(x, y)$ of the regulation map are nonempty.

We still need a definition of transversality of the rules before stating an adequate characterization of playability.

DEFINITION 2.3. We shall say that the rules P and Q are *transversal* if for all $(x, y) \in K$, for all perturbations $(e, f) \in X \times Y$, there exists (u, v) satisfying

(i) $u \in DP(y, x)(v) + e$,

(ii) $v \in DQ(x, y)(u) + f$.

We shall say that they are *strongly transversal* if

For all $(x, y) \in K$, there exist $c > 0$, $\delta > 0$ such that for all $(x', y') \in B_K((x, y), \delta)$ and all $(e, f) \in X \times Y$, there exist solutions (u, v) to

(i) $u \in DP(y', x')(v) + e$,

(ii) $v \in DQ(x', y')(u) + f$,

satisfying $\max(\|u\|, \|v\|) \leqq \max(\|e\|, \|f\|)$.

We recall that a subset K is *sleek* if the set-valued map $x \rightsquigarrow T_K(x)$ is lower semicontinuous. In this case, $T_K(x)$ is convex. A set-valued map is said to be *sleek* if its graph is sleek.

We shall now derive from Corollary 1.1 a characterization of the playability property.

THEOREM 2.1 (Playability Theorem). *Let us assume that the functions f and g are continuous, affine with respect to the strategies and have a linear growth, that the feedback maps U and V are upper semicontinuous with compact convex images and have a linear growth, and that the rules P and Q are sleek and transversal.*

Then the rules P and Q enjoy the playability property if and only if they are playable. Furthermore, the strategies $u(\cdot)$ and $v(\cdot)$ which provide playable solutions obey the following regulation law:

$$(11) \qquad \forall t \geqq 0, \quad u(t) \in C(x(t), y(t); v(t)) \quad \text{and} \quad v(t) \in D(x(t), y(t); u(t)).$$

Proof. We apply Corollary 1.1 and prove that the playability subset of the differential game is a playability domain, i.e., that for any global state $(x, y) \in K$ of the system, there exist strategies u and v such that the pair $(f(x, y, u), g(x, y, v))$ belongs to the contingent cone $T_K(x, y)$.

Since K is the intersection of the graphs of Q and P^{-1}, we need to use a sufficient condition for the contingent cone to an intersection to be equal to the intersection of the contingent cones.

The graphs of Q and P^{-1} are sleek because the rules of the game are supposed to be so. Furthermore,

$$T_{\text{Graph}(P^{-1})}(x, y) - T_{\text{Graph}(Q)}(x, y) = X \times Y$$

because the maps P and Q are transversal: For any $(e, f) \in X \times Y$, there exists (u, v) such that $(u, v-f)$ belongs to the graph of Q and $(u+e, v)$ to the graph of P^{-1}, i.e., that $(e, f) = (u+e, v) - (u, v-f)$.

Hence, by Corollary 4.3.5 of [6, p. 149], we deduce that

$$\begin{aligned} T_K(x, y) &= T_{\text{Graph}(P^{-1})}(x, y) \cap T_{\text{Graph}(Q)}(x, y) \\ &= \text{Graph}\,(DP(y, x))^{-1} \cap \text{Graph}\,(DQ(x, y)). \end{aligned}$$

Therefore, K is a viability domain if and only if the regulation map R has nonempty values, i.e., if and only if the rules of the game are playable. □

The regulation law (11) describes how the players must behave to keep the state of the system playable. A first question arises: Do the domains of the set-valued maps

(i) $C(x, y): v \rightsquigarrow C(x, y; v)$,

(ii) $D(x, y): u \rightsquigarrow D(x, y; u)$

coincide with $U(x, y)$ and $V(x, y)$, respectively?

PROPOSITION 2.1. *We posit the assumptions of Theorem* 2.1. *Let us assume that for all $(x, y) \in K$,*

$$(12) \qquad \begin{aligned} &\text{(i)} \quad \text{Dom}\,(C(x, y)) = V(x, y), \\ &\text{(ii)} \quad \text{Dom}\,(D(x, y)) = U(x, y). \end{aligned}$$

Then the rules are playable.

Proof. We deduce it from Kakutani's fixed point theorem, since the set $R(x, y)$ is the set of fixed points of the set-valued map

$$(u, v) \rightsquigarrow C(x, y; v) \times D(x, y; u)$$

defined on the convex compact subset $U(x, y) \times V(x, y)$ to itself. This set-valued map has nonempty values by assumption, which are moreover convex since the rules P and Q being sleek, the graphs of the contingent derivatives $DP(x, y)$ and $DQ(x, y)$ are

convex. They are also closed. This implies that the graph of $(u, v) \rightsquigarrow C(x, y; v) \times D(x, y; u)$ is closed. Hence we can apply Kakutani's fixed point theorem. □

3. Feedback solutions to differential games. When we know the regulation law (11), *playing the game* amounts to choosing for each pair $(x, y) \in K$ playable strategies (u, v) in the regulation set $R(x, y)$ through *playable feedbacks.*

We begin by looking for single-valued playable feedbacks $(\tilde{u}, \tilde{v})$, which are selections of the regulation map R in the sense that

$$\forall (x, y) \in K, \quad (x, y) \mapsto (\tilde{u}(x, y), \tilde{v}(x, y)) \in R(x, y)$$

or, equivalently, solutions to the system

$$\forall (x, y) \in K, \quad \begin{cases} \tilde{u}(x, y) \in C(x, y; \tilde{v}(x, y)), \\ \tilde{v}(x, y) \in D(x, y; \tilde{u}(x, y)). \end{cases}$$

For instance, continuous selections of the set-valued map R provide continuous playable feedbacks $(\tilde{u}, \tilde{v})$ such that the system of differential equations

$$\begin{aligned} x'(t) &= f(x(t), y(t), \tilde{u}(x(t), y(t))), \\ y'(t) &= g(x(t), y(t), \tilde{v}(x(t), y(t))) \end{aligned} \tag{13}$$

does have solutions which are playable.

Michael's continuous selection theorem, as well as other selection procedures we shall use, require the lower semicontinuity of the regulation map R.

Our next objective is then to provide criteria under which the regulation map is lower semicontinuous. For that purpose, we need to strengthen the concept of playable rules.

DEFINITION 3.1. We associate with any perturbation (e, f) the *retroaction rules* $C_{(e,f)}$ and $D_{(e,f)}$ defined by

$$C_{(e,f)}(x, y; v) = \{u \in U(x, y) \mid f(x, y; u) \in DP(y, x)(g(x, y, v) - f) + e\}$$

and

$$D_{(e,f)}(x, y; u) = \{v \in V(x, y) \mid g(x, y, v) \in DQ(x, y)(f(x, y; u) - e) + f\}$$

and the *regulation map* $R_{(e,f)}$ defined by

$$R_{(e,f)}(x, y) = \{(u, v) \mid u \in C_{(e,f)}(x, y; v) \text{ and } v \in D_{(e,f)}(x, y; u)\}.$$

We shall say that the rules P and Q are *strongly playable* if

For all $(x, y) \in K$, there exist $\gamma > 0$, $\delta > 0$ such that for all $(x', y') \in B_K((x, y), \delta)$ and all $(e, f) \in \gamma B$, $R_{(e,f)}(x', y') \neq \emptyset$.

THEOREM 3.1. *Let us assume that the functions f and g are continuous, affine with respect to the strategies and have a linear growth, that the feedback maps U and V are upper semicontinuous with compact convex images and have a linear growth, and that the rules P and Q are sleek, strongly transversal and strongly playable.*

Then the regulation map R is lower semicontinuous with closed convex images.

In particular, there exist continuous playable feedbacks $(\tilde{u}, \tilde{v})$.

Proof. We use the lower semicontinuity criterion of the intersection and the inverse image of lower semicontinuous set-valued maps.

First, we need to prove that the set-valued map

$$(x, y) \rightsquigarrow T_K(x, y) := \text{Graph}\,(DP(y, x)^{-1}) \cap \text{Graph}\,(DQ(x, y))$$

is lower semicontinuous. But this follows from the strong transversality of the rules P and Q and the lower semicontinuity criterion (see Theorem 1.5.5 of [6, p. 53]).

We observe that $U \times V$ being upper semicontinuous with compact values, it maps a neighborhood of each point to a compact set. Since we can write

$$R(x, y) = \{(u, v) \in U(x, y) \in V(x, y) \mid (f(x, y; u)g(x, y; v)) \in T_K(x, y)\}$$

and since both $U \times V$ and T are lower semicontinuous with convex images, strong playability of the retroaction rules implies that the regulation map R is lower semicontinuous. □

Unfortunately, the proof of Michael's continuous selection theorem is not constructive. We would rather trade the continuity of the playable control with some explicit and computable property, such as $u^0(x, y)$ being the element of minimal norm in $R(x, y)$, or other properties. Hence we need to prove the existence of a solution to the differential equation (13) for such noncontinuous feedbacks.

We shall provide a general method of construction of such playable feedbacks. For that purpose it is useful to introduce the following definition.

DEFINITION 3.2 (selection procedure). A selection procedure of the regulation map $R: X \times Y \rightsquigarrow U \times V$ is a set-valued map $S_R: X \times Y \rightsquigarrow U \times V$

(i) For all $(x, y) \in K$, $S(R(x, y)) := S_R(x, y) \cap R(x, y) \neq \emptyset$,

(ii) The graph of S_R is closed,

and the valued map $S(R): (x, y) \rightsquigarrow S(R(x, y))$ is called the *selection* of R.

It is said to be *convex-valued* if its values are convex and *single-valued* if, moreover,

$$\forall (x, y) \in \text{Dom}(R), \quad S_R(x, y) \cap R(x, y) = \{\tilde{u}(x, y), \tilde{v}(x, y)\} \tag{14}$$

is a singleton.

THEOREM 3.2. *We posit the assumptions of Theorem* 3.1 *and we suppose that K is a playability domain.*

Let S_R be a convex-valued selection procedure of the regulation map R. Then, for any initial state $(x_0, y_0) \in K$, there exists a playable solution starting at (x_0, y_0) to the differential inclusion

(i) $x'(t) = f(x(t), y(t); u(t))$,

(ii) $y'(t) = g(x(t), y(t); v(t))$,

(iii) *for almost all t,* $(u(t), v(t)) \in S(R(x(t), y(t)))$.

In particular, if the selection procedure is single-valued, then the strategies

$$(\tilde{u}(x, y), \tilde{v}(x, y)) \quad \textit{defined by } (14)$$

are single-valued playable feedback controls.

Proof. Since the convex selection procedure S_R has a closed graph and convex values, we can replace the differential game (1) by the controlled system

$$\begin{aligned} &\text{(i)} \quad x'(t) = f(x(t), y(t); u(t)), \\ &\text{(ii)} \quad y'(t) = g(x(t), y(t); v(t)), \\ &\text{(iii)} \quad \text{for almost all } t, \end{aligned} \tag{15}$$

$$(u(t), v(t)) \in (U(x(t), y(t)) \times V(x(t), y(t))) \cap S_R(x(t), y(t)),$$

which satisfies the assumptions of the Viability Theorem. It remains to check that K is still a viability domain for this "smaller system."

But by construction, we know that for all $(x, y) \in K$, there exists $(u, v) \in S(R(x, y))$, which belongs to the intersection $U(x, y) \times V(x, y) \cap S_R(x, y)$ and which is such that $(f(x, y; u), g(x, y; v))$ belongs to $T_K(x)$.

Hence the new controlled system (15) enjoys the viability property, so that, for any initial state $(x_0, y_0) \in K$, there exist a viable solution and a viable control to the controlled system (15) which, for almost all $t \geqq 0$, are related by

(i) $(u(t), v(t)) \in (U(x(t), y(t)) \times V(x(t), y(t))) \cap S_R(x(t), y(t))$,

(ii) $(f(x(t), y(t); u(t)), g(x(t), y(t); v(t)) \in T_K(x(t), y(t))$.

Therefore, for almost all $t \geqq 0$, $(u(t), v(t))$ belongs to the intersection of $R(x(t), y(t))$ and $S_R(x(t), y(t))$, i.e., to the selection $S(R(x(t), y(t)))$ of the regulation map R. □

We can now multiply the possible corollaries, by giving several instances of selection procedures of set-valued maps.

We begin by cooperative procedures, where the players agree on criteria $\sigma(x, y; \cdot, \cdot)$ for selecting strategies in the regulation sets $R(x, y)$.

Example. Cooperative behavior. Let $\sigma : \text{Graph}(R) \mapsto \mathbf{R}$ be continuous.

PROPOSITION 3.1. *We posit the assumptions of Theorem* 3.1. *Let σ be continuous on* Graph (R) *and convex with respect to the pair (u, v). Then, for any initial state $(x_0, y_0) \in K$, there exist a playable solution starting at (x_0, y_0) and playable strategies to the differential game* (1) *which are regulated by*

for almost all $t \geqq 0$,

$$\sigma(x(t), y(t); u(t), v(t)) = \inf_{u', v' \in R(x(t), y(t))} \sigma(x(t), y(t); u', v').$$

In particular, the game can be played by the slow feedbacks of minimal norm:

$$(u^0(x, y), v^0(x, y)) \in R(x, y),$$

$$\|(u^0(x, y)\|^2 + \|v^0(x, y)\|^2) = \min_{(u,v) \in R(x,y)} (\|u\|^2 + \|v\|^2).$$

Proof. We introduce the set-valued map S_R defined by

$$S_R(x, y) := \left\{ (u, v) \in Y \,|\, \sigma(x, y; u, v) \leqq \inf_{(u', v') \in R(x,y)} \sigma(x, y; u', v') \right\}.$$

It is a convex *selection procedure* of R. Indeed, since R is lower semicontinuous, the function

$$(x, y; u, v) \mapsto \sigma(x, y; u, v) + \sup_{(u', v') \in R(x,y)} (-\sigma(x, y; u', v'))$$

is lower semicontinuous thanks to the Maximum Theorem. Then the graph of S_R is closed because

$$\text{Graph}(S_R) = \left\{ (x, y) \,|\, \sigma(x, y; u, v) + \sup_{(u', v') \in R(x,y)} (-\sigma(x, y; u', v')) \leqq 0 \right\}.$$

The images are obviously convex. Consequently, the graph of R also being closed, so is the selection $S(R)$ equal to

$$S(R(x, y)) = \left\{ (u, v) \in R(x, y) \,|\, \sigma(x, y; u, v) \leqq \inf_{(u', v') \in R(x,y)} (\sigma(x, y; u', v')) \right\}.$$

We then apply Theorem 3.2. We observe that when we take

$$\sigma(x, y; u, v) := \|u\|^2 + \|v\|^2,$$

the selection procedure is single-valued and yields the elements of minimal norm. □

Example. Noncooperative behavior. We can also choose strategies in the regulation sets $R(x, y)$ in a noncooperative way, as saddle points of a function $a(x, y; \cdot, \cdot)$.

PROPOSITION 3.2. *We posit the assumptions of Theorem* 3.1 *and we suppose that* K *is a playability domain. Let us assume that* $a: X \times Y \times U \times V \to \mathbf{R}$ *satisfies*

(i) *a is continuous,*

(ii) *For all* $(x, y, v) \in X \times V$, $u \mapsto a(x, y; u, v)$ *is convex,*

(iii) *For all* $(x, y; u) \in X \times U$, $v \mapsto a(x, y; u, v)$ *is concave.*

Then, for any initial state $(x_0, y_0) \in K$, *there exist a playable solution starting at* (x_0, y_0) *and playable strategies to the differential game* (1) *which are regulated by*

$$\text{for almost all } t \geqq 0, \begin{cases} \text{(i)} & (u(t), v(t)) \in R(x(t), y(t)), \\ \text{(ii)} & \text{For all } (u', v') \in R(x(t), y(t)), \\ & a(x(t), y(t); u(t), v') \leqq a(x(t), y(t); u(t), v(t)) \\ & \qquad \leqq a(x(t), y(t); u', v(t)). \end{cases}$$

Proof. We prove that the set-valued map S_R associating with any $(x, y) \in K$ the subset

$$S_R(x, y) := \{(u, v) \in U \times V \text{ such that for all } (u', v') \in R(x, y),\ a(x, u, v') \leqq a(x, u', v)\}$$

is a convex selection procedure of R. The associated selection map $S(R(\cdot))$ associates with any $x \in X$ the subset

$$S(R(x, y)) := \{(u, v) \in R(x, y) \text{ such that for all } (u', v') \in R(x, y),\ a(x, y; u, v') \leqq a(x, y; u', v)\}$$

of saddle points of $a(x, y; \cdot, \cdot)$ in $R(x, y)$. Von Neumann's minimax theorem states that the subsets $S(R(x, y))$ of saddle points are not empty since $R(x, y)$ are convex and compact. The graph of S_R is closed thanks to the assumptions and the Maximum Theorem because it is equal to the lower section of a lower semicontinuous function:

$$\text{Graph}(S_R) = \left\{ (x, y) \,\middle|\, \sup_{(u', v') \in R(x, y)} (a(x, y; u, v') - a(x, y; u', v)) \leqq 0 \right\}.$$

We then apply Theorem 3.2. □

4. Discriminating and leading feedbacks. We now address the question of finding criteria for the playability domain K to be Xavier's discriminating domain, and for finding Xavier's feedback strategies which are selections of the set-valued map $(x, y, v) \rightsquigarrow A(x, y, v) \subset U(x, y)$ defined by

$$A(x, y; v) := \{u \in U(x, y) \,|\, (u, v) \in R(x, y)\}.$$

Such feedbacks are called *discriminating feedbacks.* If we assume that Xavier has access to the strategies chosen by Yvette, he can keep the states of the system playable by "playing" a discriminating control whatever the choice of Yvette through a discriminating feedback.

Then, we shall investigate whether we can find (possibly, single-valued) selections of such a set-valued map A, and for that, provide sufficient conditions for A to be lower semicontinuous.

We first observe that A can be written in the form

$$A(x, y; v) := C(x, y; v) \cap (D(x, y))^{-1}(v).$$

The first assumption we must make for obtaining discriminating feedbacks for Yvette is that the domain of the set-valued maps $A(x, y; \cdot)$ are not empty, i.e., that

For all $v \in V(x, y)$, there exists $u \in U(x, y)$ such that
$f(x, y; u) \in DP(y, x)(g(x, y; v)) \cap DQ(x, y)^{-1}(g(x, y; v))$.

We shall actually strengthen it a bit to get the lower semicontinuity of A, by assuming that

(16) For all $(x, y) \in K$ and all $v \in V(x, y)$, there exist $\delta > 0$, and $\gamma > 0$ such that for all $(x', y') \in B_K(x, y, \delta)$, $v' \in B(v, \delta) \cap V(x', y')$, and all $\|e_i\| \leqq \gamma$ $(i = 1, 2)$ there exist $u \in U(x', y')$ such that $f(x', y'; u) \in (DP(y', x')(g(x', y'; v')) - e_1) \cap (DQ(x', y')^{-1}(g(x', y'; v')) - e_2)$.

PROPOSITION 4.1. *We posit the assumptions of Theorem* 3.1, *where we replace strong playability by assumption* (16), *and we assume further that the norms of the closed convex processes* $DP(y, x)$ *and* $DQ(x, y)^{-1}$ *are bounded. Then the set-valued map* A *is lower semicontinuous.*

Proof. First, we must prove that C is lower semicontinuous, and, for that purpose, that $(x, y, w) \rightsquigarrow DP(y, x)(w)$ is lower semicontinuous.

By a generalization of the Banach-Steinhauss theorem to closed convex process (see Theorem 2.3.2 of [6, p. 61]), we know that it is sufficient to prove that

$$(x, y) \rightsquigarrow \text{Graph}\,(DP(y, x)) \text{ is lower semicontinuous}$$

and that

$$\|DP(y, x)\| := \sup_{\|w\| \leqq 1} \inf_{u \in DP(y,x)(w)} \|u\| < +\infty.$$

This is the case because P is assumed to be sleek and because we have assumed that the norms of the derivatives are bounded. Therefore, the set-valued map

$$(x, y, v) \rightsquigarrow DP(y, x)(g(x, y; v))$$

is also lower semicontinuous.

The lower semicontinuity criterion and assumption (16) imply that $(x, y, v) \rightsquigarrow C(x, y; v)$ is lower semicontinuous.

The same proof shows that the set-valued map $(x, y, v) \rightsquigarrow DQ(x, y)^{-1}(v)$ is also lower semicontinuous. Since A is the intersection of these two set-valued maps, we again apply the lower semicontinuity criterion to deduce that A is lower semicontinuous, which is possible thanks to assumption (16). □

THEOREM 4.1. *We posit the assumptions of Proposition* 4.1. *For any continuous feedback control* $(x, y) \mapsto \tilde{v}(x, y)$ *played by Yvette, there exists a continuous single-valued feedback* $\tilde{u}(x, y)$ *played by Xavier such that the differential equation* (13) *has playable solutions for any initial state* $(x_0, y_0) \in K$.

More generally, let S_A *be a convex selection procedure of the set-valued map* A. *Then, for any continuous feedback control* $(x, y) \mapsto \tilde{v}(x, y)$ *played by Yvette, for any initial state* $(x_0, y_0) \in K$, *there exists a playable solution starting at* (x_0, y_0) *to the differential game*

(i) $x'(t) = f(x(t), y(t); u(t))$,

(ii) $y'(t) = g(x(t), y(t); \tilde{v}(x(t), y(t)))$,

(iii) $u(t) \in S(A(x(t), y(t); \tilde{v}(x(t), y(t))))$.

In particular, if the selection procedure is single-valued, then the control $\tilde{u}(x, y)$ *defined by*

$$\tilde{u}_{\tilde{v}}(x, y) := S(A(x, y; \tilde{v}(x, y)))$$

is a single-valued feedback control.

This is the case, for instance, when Xavier plays the feedback control $u^0_{\tilde{v}}(x, y)$ *of minimal norm in the set* $A(x, y; \tilde{v}(x, y))$.

Proof. Whenever Yvette plays a continuous feedback $\tilde{v}(x, y)$, K remains a playability domain for the system

(i) $x'(t) = f(x(t), y(t); u(t))$,

(ii) $y'(t) = g(x(t), y(t); \tilde{v}(x(t), y(t)))$,

(iii) $u(t) \in A(x(t), y(t); \tilde{v}(x(t), y(t)))$.

Since the set-valued map $(x, y) \rightsquigarrow A(x, y; \tilde{v}(x, y))$ is lower semicontinuous, it contains continuous selections $\tilde{u}(x, y)$ which therefore yield playable selections.

We can also use more constructive convex selection procedures of the set-valued map $(x, y) \rightsquigarrow A(x, y; \tilde{v}(x, y))$ and deduce that Xavier can implement playable solutions by playing strategies $u(t)$ in the selection $S(A(x(t), y(t); \tilde{v}(x(t), y(t))))$. □

A much better situation for Xavier occurs when he can find feedback strategies $\tilde{u}$ which are selections of the set-valued map B defined by

$$B(x, y) := \bigcap_{v \in V(x,y)} A(x, y; v).$$

In other words, such a feedback allows him to implement playable solutions whatever the control $v \in V(x, y)$ chosen by Yvette, since in this case the pair (u, v) belongs to the regulation set $R(x, y)$ for any v. Such feedbacks are called *pure feedbacks.*

In order to obtain continuous single-valued feedbacks, we need to prove the lower semicontinuity of the set-valued map B, which is an infinite intersection of lower semicontinuous set-valued maps.

THEOREM 4.2. *We posit the assumptions of Proposition* 4.1. *We assume further that there exist positive constants δ and γ such that for all $(x', y') \in B_K((x, y), \delta)$, we have*

(17) *For all $v \in V(x', y')$ and all $e_v^i \in \gamma B$, $(i = 1, 2)$, there exists $u \in U(x', y')$ such that $f(x', y'; u) \in DP(y', x'; v) + e_v^1$ and $g(x', y'; v) \in DQ(x', y'; u) + e_v^2$.*

Then the set-valued map B is lower semicontinuous and there exist continuous single-valued pure feedback strategies for Xavier.

Proof. We observe that V is upper semicontinuous with compact values, that A is lower semicontinuous and has its images in a fixed compact set, and that assumption (17) implies obviously that there exist positive constants δ and γ such that for all $(x', y') \in B_K((x, y), \delta)$, we have

$$\forall v \in V(x', y'), \quad \forall e_v \in \gamma B, \quad \bigcap_{v \in V(x', y')} (A(x', y'; v) - e_v) \neq \emptyset.$$

This theorem follows then from the general criterion on the lower semicontinuity of an infinite intersection of lower semicontinuous set-valued maps.

THEOREM 4.3. *Let us consider a metric space X, normed vector-spaces Y and Z, and set-valued maps $F : X \times Y \rightsquigarrow Z$ and $H : X \rightsquigarrow Y$. We assume that*

(i) *F is lower semicontinuous with convex values,*

(ii) *H is upper semicontinuous with compact values,*

and that there exist positive constants γ, δ, c such that for every single-valued map $e : Y \mapsto \gamma B$ we have

(18) $$\forall x' \in B(x, \delta), \quad cB \cap \bigcap_{y \in H(x')} (F(x', y) - e(y)) \neq \emptyset.$$

Then the set-valued map $G : X \rightsquigarrow Z$ defined by

$$\forall x \in X, \quad G(x) := \bigcap_{y \in H(x)} F(x, y)$$

is lower semicontinuous (with nonempty convex images).

Remark. When the set-valued map F is locally bounded (in the sense that it maps some neighborhood of each point to a bounded subset), we do not need the constant c and we can replace (18) by

$$\forall x' \in B(x, \delta), \quad \bigcap_{y \in H(x')} (F(x', y) - e(y)) \neq \emptyset. \qquad \square$$

Proof. Let us choose any sequence of elements $x_n \in \mathrm{Dom}\,(F)$ converging to x and $z \in G(x)$. We must approximate z by elements $z_n \in G(x_n)$.

We introduce the following numbers:

$$e_n := \sup_{y \in H(x_n)} d(z, F(x_n, y))/2. \tag{19}$$

Now, let us choose for each $y \in H(x_n)$ an element $u_n(y) \in F(x_n, y)$ satisfying

$$\|z - u_n(y)\| \leqq 2d(z, F(x_n, y)) \leqq e_n$$

and set $\theta_n := \gamma/(\gamma + e_n)$. Consequently,

$$\theta_n(z - u_n(y)) \in \theta_n e_n B = (1 - \theta_n)\gamma B,$$

so that there exists $a_n(y) \in \gamma B$ such that

$$\theta_n(z - u_n(y)) = (1 - \theta_n)a_n(y).$$

Therefore, assumption (18) implies the existence for all n large enough of elements $w_n \in cB$ and elements $v_n(y) \in F(x_n, y)$ such that $a_n(y) = v_n(y) - w_n$ for all $y \in H(x_n)$.

Hence we can write

$$\theta_n(z - u_n(y)) = (1 - \theta_n)(v_n(y) - w_n)$$

so that the common value

$$z_n := \theta_n z + (1 - \theta_n)w_n = \theta_n u_n(y) + (1 - \theta_n)v_n(y)$$

does not depend on y, belongs to all $F(x_n, y)$ (by convexity), and converges to z because

$$\|z - z_n\| = (1 - \theta_n)\|z - w_n\| \leqq (1 - \theta_n)(\|z\| + c)$$

and because $1 - \theta_n = e_n/(\gamma + e_n)$ converges to zero for e_n converges to zero thanks to the following lemma. $\square$

LEMMA 4.1. *Let us assume that F is lower semicontinuous and that H is upper semicontinuous with compact images. Then the numbers e_n defined by* (19) *converge to zero.*

Proof. Since F is lower semicontinuous [6, Cor. 1.4.17, p. 49] the Maximum Theorem implies that the function

$$(x, y, z) \mapsto d(z, F(x, y))$$

is upper semicontinuous. Therefore, for any $\varepsilon > 0$ and any $y \in H(x)$, there exist an integer N_y and a neighborhood $\mathcal{V}_y$ of y such that

$$\forall y' \in \mathcal{V}_y, \quad \forall n \geqq N_y, \quad d(z, F(x_n, y')) \leqq \varepsilon \tag{20}$$

because $d(z, F(x, y)) = 0$. Hence the compact set $H(x)$ can be covered by p neighborhoods $\mathcal{V}_{y_i}$. Furthermore, H being upper semicontinuous, there exists an integer N_0 such that

$$\forall n \geqq N_0, \quad H(x_n) \subset \bigcup_{i=1,\cdots,p} \mathcal{V}_{y_i}.$$

Set $N := \max_{i=0,\cdots,p} N_{y_i}$. Then, for all $n \geqq N$ and $y \in H(x_n)$, y belongs to some $\mathcal{V}_{y_i}$, so that, by (20), $d(z, F(x_n, y)) \leqq \varepsilon$. Thus,

$$\forall n \geqq N, \quad e_n := \sup_{y \in H(x_n)} d(z, F(x_n, y))/2 \leqq \frac{\varepsilon}{2},$$

i.e., our lemma is proved. □

5. Closed-loop decision rules. Actually, although differential games can be played through retroaction rules, there are many games where players *act on the velocities of the strategies* rather than on the state of the controls. We can regard changes of strategies as *decisions* of players.

This leads us to introduce the following definition. We shall call *decisions* the derivatives of the strategies.

Then, in order to deal with decisions defined in such a sense, we must now assume that players use open-loop strategies $u(\cdot)$ and $v(\cdot)$ which are *absolutely continuous* and obey a growth condition of the type[4]

$$(21) \qquad \begin{aligned} &\text{(i)} \quad \|u'(t)\| \leqq \rho(\|u(t)\| + 1), \\ &\text{(ii)} \quad \|v'(t)\| \leqq \sigma(\|v(t)\| + 1). \end{aligned}$$

We shall refer to them as "smooth open-loop controls," the nonnegative parameters ρ and σ being fixed once and for all. We denote by $\mathcal{K}$ the subset

$$(z, u, v) \in \mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q \text{ such that } u \in U(z) \text{ and } v \in V(z).$$

Instead of finding largest playability or winability domains in the state space, we shall look for analogous concepts in the state-strategy space. We shall determine set-valued maps which allow players to win in the sense that either

$$(22) \qquad \forall t \geqq 0, \quad u(t) \in U(z(t))$$

or

$$(23) \qquad \forall t \geqq 0, \quad v(t) \in V(z(t))$$

or both. Roughly speaking, Xavier may win as long as its opponent allows him to choose at each instant $t \geqq 0$ strategies $u(t)$ in the subset $U(z(t))$, and must lose if for any choice of open-loop controls, there exists a time $T > 0$ such that $u(T) \notin U(z(T))$.

DEFINITION 5.1. Let (u_0, v_0, z_0) be an initial situation such that initial strategies $u_0 \in U(z_0)$ and $v_0 \in V(z_0)$ of the two players are consistent with the initial state z_0.

We shall say that

— Xavier *must win* if and only if for all smooth open-loop strategies $u(\cdot)$ and $v(\cdot)$ starting at u_0 and v_0, there exists a solution $z(\cdot)$ to (3) and (21) starting at z_0 such that (22) is satisfied.

— Xavier *may win* if and only if there exist smooth open-loop strategies $u(\cdot)$ and $v(\cdot)$ starting at u_0 and v_0 and a solution $z(\cdot)$ to (3) and (21) starting at z_0 such that (22) is satisfied.

— Xavier *must lose* if and only if for any smooth open-loop strategy $u(\cdot)$ and $v(\cdot)$ starting at u_0 and v_0 and solution $z(\cdot)$ to (3) and (21) starting at z_0, there exists a time $T > 0$ such that

$$u(T) \notin U(z(T)).$$

[4] We can replace $\rho(\|u\| + 1)$ by any continuous function $\varphi(u)$ with linear growth.

— The initial situation is *stable* if and only if there exist open-loop strategies $u(\cdot)$ and $v(\cdot)$ starting at u_0 and v_0 and a solution $z(\cdot)$ to (3) and (21) starting at z_0 satisfying *both* relations (22) and (23).

Naturally, if both Xavier and Yvette must win, then both relations (22) and (23) are satisfied. This is not necessarily the case when both Xavier and Yvette may win, and this is why we need to introduce the concept of stability.

THEOREM 5.1. *Let us assume that h is continuous with linear growth and that the graphs of U and V are closed. Let the growth rates ρ and σ be fixed.*

There exist five (possibly empty) closed set-valued feedback maps from $\mathbf{R}^n$ to $\mathbf{R}^p \times \mathbf{R}^q$ having the following properties:

— *$R_U \subset U$ is such that whenever $(u_0, v_0) \in R_U(z_0)$, Xavier may win and that whenever $(u_0, v_0) \notin R_U(z_0)$, Xavier must loose.*

— *If h is Lipschitz, $S_U \subset R_U$ is the largest closed set-valued map such that whenever $(u_0, v_0) \in S_U(z_0)$, Xavier must win.*

— *$S_V \subset R_V \subset V$, which have analogous properties.*

— *$R_{UV} \subset R_U \cap R_V$ is the largest closed set-valued map such that any initial situation satisfying $(u_0, v_0) \in R_{UV}(z_0)$ is stable.*

Knowing these five set-valued feedback maps, we can split the domain $\mathcal{K}$ of initial situations into ten areas which describe the behavior of the differential game from the position of the initial situation.

In particular, the complement of the graph of R_{UV} in the intersection of the graphs of R_U and R_V is the instability region, where either Xavier or Yvette may win, but not both together.

TABLE 1
The 10 areas of the domain of the differential game.

$(z_0, u_0, v_0) \in$	Graph (S_U)	Graph (R_U)	$\mathcal{K} \backslash$ Graph (R_U)
Graph (S_V)	Xavier must win Yvette must win	Xavier may win Yvette must win	Xavier must loose Yvette must win
Graph (R_V)	Xavier must win Yvette may win	? ? ? ? STABILITY ? ? ? ?	Xavier must loose Yvette may win
$\mathcal{K} \backslash$ Graph (R_V)	Xavier must win Yvette must loose	Xavier may win Yvette must loose	Xavier must loose Yvette must loose

The problem is to characterize these five set-valued maps, the existence of which is now guaranteed, by solving the "contingent extension" of the partial differential equation[5]

$$\frac{\partial \Phi}{\partial z} \cdot h(z, u, v) - \rho(\|u\| + 1) \left\| \frac{\partial \Phi}{\partial u} \right\| - \sigma(\|v\| + 1) \left\| \frac{\partial \Phi}{\partial v} \right\| \leqq 0, \tag{24}$$

[5] If Φ is a solution to this partial differential equation, we can check that for any initial situation $(z_0, u_0, v_0) \in \mathrm{Dom}(\Phi)$, there exists a smooth solution $(z(\cdot), u(\cdot), v(\cdot))$ such that

$$t \to \Phi(z(t), u(t), v(t)) \quad \text{is nonincreasing.}$$

This property remains true for the solutions to the contingent partial differential equation (27).

which can be written in the following way:

$$\frac{\partial \Phi}{\partial z} \cdot h(z, u, v) + \inf_{\|u'\| \leqq \rho(\|u\|+1)} \frac{\partial \Phi}{\partial u} \cdot u' + \inf_{\|v'\| \leqq \sigma(\|v\|+1)} \frac{\partial \Phi}{\partial v} \cdot v' \leqq 0.$$

We shall also introduce the partial differential equation[6]

$$\frac{\partial \Phi}{\partial z} \cdot h(z, u, v) + \rho(\|u\|+1) \left\| \frac{\partial \Phi}{\partial u} \right\| + \sigma(\|v\|+1) \left\| \frac{\partial \Phi}{\partial v} \right\| \leqq 0, \tag{25}$$

which can be written in the following way:

$$\frac{\partial \Phi}{\partial z} \cdot h(z, u, v) + \sup_{\|u'\| \leqq \rho(\|u\|+1)} \frac{\partial \Phi}{\partial u} \cdot u' + \sup_{\|v'\| \leqq \sigma(\|v\|+1)} \frac{\partial \Phi}{\partial v} \cdot v' \leqq 0.$$

The link between the feedback maps and the solutions to the solutions to these partial differential equations is provided by the indicators of the graphs: we associate with the set-valued maps S_U, R_U, and R_{UV} the functions Φ_U, Ψ_U, and Ψ_{UV} from $\mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q$ to $\mathbf{R}_+ \cup \{+\infty\}$ defined by

$$\text{(i)} \quad \Phi_U(z, u, v) := \begin{cases} 0 & \text{if } (u, v) \in S_U(z), \\ +\infty & \text{if } (u, v) \notin S_U(z), \end{cases} \tag{26}$$

$$\text{(ii)} \quad \Psi_U(z, u, v) := \begin{cases} 0 & \text{if } (u, v) \in R_U(z), \\ +\infty & \text{if } (u, v) \notin R_U(z), \end{cases}$$

$$\text{(iii)} \quad \Psi_{UV}(z, u, v) := \begin{cases} 0 & \text{if } (u, v) \in R_{UV}(z), \\ +\infty & \text{if } (u, v) \notin R_{UV}(z), \end{cases}$$

and the functions Ψ_V and Φ_V associated with the set-valued map R_V and S_V in an analogous way.

These functions being only lower semicontinuous, but not differentiable, cannot be solutions to either partial differential equations (24) or (25). But we can use the *contingent epiderivatives* of any function $\Phi: \mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q \to \mathbf{R} \cup \{+\infty\}$ and replace the partial differential equations (24) and (25) by the contingent partial differential equations

$$\inf_{\substack{\|u'\| \leqq \rho(\|u\|+1) \\ \|v'\| \leqq \sigma(\|v\|+1)}} D_\uparrow \Phi(z, u, v)(h(z, u, v), u', v') \leqq 0 \tag{27}$$

and

$$\sup_{\substack{\|u'\| \leqq \rho(\|u\|+1) \\ \|v'\| \leqq \sigma(\|v\|+1)}} D_\uparrow \Phi(z, u, v)(h(z, u, v), u', v') \leqq 0, \tag{28}$$

respectively.

Let Ω_U and Ω_V be the indicators of the graphs of the set-valued maps U and V defined by

$$\text{(i)} \quad \Omega_U(z, u, v) := \begin{cases} 0 & \text{if } u \in U(z), \\ +\infty & \text{if } u \notin U(z), \end{cases}$$

$$\text{(ii)} \quad \Omega_V(z, u, v) := \begin{cases} 0 & \text{if } v \in V(z), \\ +\infty & \text{if } v \notin V(z). \end{cases}$$

[6] We can check that if f is Lipschitz and Φ is a solution to this partial differential equation, for any initial situation $(z_0, u_0, v_0) \in \mathrm{Dom}(\Phi)$, any smooth solution $(z(\cdot), u(\cdot), v(\cdot))$ satisfies that

$$t \to \Phi(z(t), u(t), v(t)) \quad \text{is nonincreasing.}$$

This property remains true for the solutions to the contingent partial differential equation (28).

THEOREM 5.2. *We posit the assumptions of Theorem* 5.1. *Then*

— Ψ_U *is the smallest lower semicontinuous solution to the contingent partial differential equation* (27) *larger than or equal to* Ω_U.

— Ψ_V *is the smallest lower semicontinuous solution to the contingent partial differential equation* (27) *larger than or equal to* Ω_V.

— Ψ_{UV} *is the smallest lower semicontinuous solution to the contingent partial differential equation* (27) *larger than or equal to* $\max(\Omega_U, \Omega_V)$.

— *If* h *is Lipschitz,* Φ_U *is the smallest lower semicontinuous solution to the contingent partial differential equation* (28) *larger than or equal to* Ω_U.

— *If* h *is Lipschitz,* Φ_V *is the smallest lower semicontinuous solution to the contingent partial differential equation* (28) *larger than or equal to* Ω_V.

If any of the above solutions is the constant $+\infty$, the corresponding feedback map is empty.

Proof of Theorem 5.1. Let us denote by B the unit ball and introduce the set-valued map F defined by

$$H(z, u, v) := \{h(z, u, v)\} \times \rho(\|u\|+1)B \times \sigma(\|v\|+1)B.$$

The evolution of the differential game described by equations (3) and (21) is governed by the differential inclusion

$$(z'(t), u'(t), v'(t)) \in H(z(t), u(t), v(t)).$$

— Since the graph of U is closed, we know that there exists a largest closed viability domain contained in Graph $(U) \times \mathbf{R}^q$, which is the set of initial situations (z_0, u_0, v_0) such that there exists a solution $(z(\cdot), u(\cdot), v(\cdot))$ to this differential inclusion remaining in this closed set. This is the graph of R_U. Indeed, if $(u_0, v_0) \in R_U(z_0)$, there exists a solution to the differential inclusion remaining in the graph of U, i.e., Xavier may win. If not, all solutions starting at (z_0, u_0, v_0) must leave this domain in finite time.

The set-valued feedback map R_V is defined in an analogous way.

— For the same reasons, the graph of the set-valued feedback map R_{UV} is the largest closed viability domain of the set $\mathcal{K}$ of initial situations.

— When h is Lipschitz, so is F. Then the solution-map $S(z_0, u_0, v_0)$ is also Lipschitz thanks to Filippov's theorem,[7] so that the subset of initial situations such that all the functions of $S(z_0, u_0, v_0)$ remain in a closed subset is also closed. This is the largest closed invariant domain by F of this closed subset. Then the largest closed invariant domain contained in Graph $(U) \times \mathbf{R}^q$ is the graph of the set-valued feedback map S_U. □

Proof of Theorem 5.2. We recall that thanks to Haddad's viability theorem, a subset $L \subset \mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q$ is a viability domain of F if and only if

$$\forall (z, u, v) \in L, \quad T_L(z, u, v) \cap H(z, u, v) \neq \emptyset.$$

Let Ψ_L denote the indicator of L. We know that the Viability Theorem can be reformulated in the following way.

L is a closed viability domain if and only if its indicator function Ψ_L is a solution to the contingent partial differential equation (27).

— Hence to say that the graph of R_U is the largest closed viability domain contained in the graph of U amounts to saying that its indicator Ψ_U is the smallest lower semicontinuous solution to the contingent partial differential equation (27) larger

[7] See [6, p. 402].

than or equal to the indicator Ω_U of Graph $(U) \times \mathbf{R}^q$. The same reasoning shows that indicator Ψ_V of R_V is the smallest lower semicontinuous solution to the contingent partial differential equation (27) larger than or equal to Ω_V and that the indicator Ψ_{UV} of the graph of R_{UV} is the smallest lower semicontinuous solution to the contingent partial differential equation (27) larger than or equal to the indicator of $\mathcal{K}$, which is equal to $\max(\Omega_U, \Omega_V)$.

— We know that a closed subset $L \subset \mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q$ is "invariant" by a Lipschitz set-valued map F if and only if

$$\forall (z, u, v) \in L, \quad T_L(z, u, v) \subset H(z, u, v).$$

This condition can be reformulated in terms of contingent epiderivative of the indicator function Ψ_L of L saying that

$$\forall (z, u, v) \in L, \quad \sup_{w \in H(z,u,v)} D_\uparrow \Psi_L(z, u, v)(w) = 0.$$

Hence to say that the graph of S_U is the largest closed invariance domain contained in the graph of U amounts to saying that its indicator $\Phi_{U'}$ is the smallest lower semicontinuous solution to the contingent partial differential equation (28) larger than or equal to the indicator Ω_U of Graph $(U) \times \mathbf{R}^q$. □

Let us denote by R one of the feedback maps R_U, R_V, R_{UV} and assume that the initial situation belongs to the graph of the set-valued feedback map R (when it is not empty). The theorem states only that there exists at least a solution $(z(\cdot), u(\cdot), v(\cdot))$ to the differential game such that

$$\forall t \geqq 0, \quad (u(t), v(t)) \in R(z(t)).$$

To implement strategy, players *must make decisions, i.e., choose velocities of controls* in an adequate way.

We observe these stable solutions.

PROPOSITION 5.1. *The solutions to the game satisfying*

$$\forall t \geqq 0, \quad (u(t), v(t)) \in R(z(t))$$

are the solutions to the system of differential inclusions

$$\begin{aligned} &\text{(i)} \quad z'(t) = h(z(t), u(t), v(t)), \\ &\text{(ii)} \quad (u'(t), v'(t)) \in G_R(z(t), u(t), v(t)), \end{aligned} \tag{29}$$

where we have denoted by G_R the R-decision map defined by

$$G_R(z, u, v) := DR_R(z, u, v)(h(z, u, v)).$$

For simplicity, we shall set $G := G_R$ whenever there is no ambiguity.

Proof. Indeed, since the function $(z(\cdot), u(\cdot), v(\cdot))$ takes its values into Graph (R) and is absolutely continuous, then its derivative $(z'(\cdot), u'(\cdot), v'(\cdot))$ belongs almost everywhere to the contingent cone

$$T_{\text{Graph}(R)}(z(t), u(t), v(t)) := \text{Graph}\,(DR(z(t), u(t), v(t))).$$

We then replace $z'(t)$ by $h(z(t), u(t), v(t))$.

The converse holds true because equation (29) makes sense only if $(z(t), u(t), v(t))$ belongs to the graph of R. □

The question arises of whether we can construct selection procedures of the decision components of this system of differential inclusions. It is convenient for this purpose to introduce the following definition.

DEFINITION 5.2 (Closed Loop Decision Rules). We say that a selection $(\tilde{c}, \tilde{d})$ of the contingent derivative of the smooth regulation map R in the direction h defined by

$$(30) \qquad \forall (z, u, v) \in \text{Graph}\,(R), (\tilde{c}(z, u, v), \tilde{d}(z, u, v)) \in DR(z, u, v)(h(z, u, v))$$

is a *closed-loop decision rule.*

The system of differential equations

$$(31) \qquad \begin{aligned} &\text{(i)} \quad z'(t) = h(z(t), u(t), v(t)), \\ &\text{(ii)} \quad u'(t) = c(z(t), u(t), v(t)), \\ &\text{(iii)} \quad v'(t) = d(z(t), u(t), v(t)) \end{aligned}$$

is called the associated *closed-loop decision game.*

Therefore, closed-loop decision rules being given for each player, the closed-loop decision system is just a system of ordinary differential equations.

It has solutions whenever the maps c and d are continuous (and if such is the case, they will be continuously differentiable).

But they also may exist when c or d or both are no longer continuous. This is the case when the decision map is lower semicontinuous thanks to Michael's theorem.

THEOREM 5.3. *Let us assume that the decision map $G := G_R$ is lower semicontinuous with nonempty closed convex values on the graph of R. Then there exist continuous decision rules c and d, so that the decision system* (31) *has a solution whenever the initial situation* $(u_0, v_0) \in R(z_0)$.

By using selection procedures introduced above, we can obtain explicit decision rules which are not necessarily continuous, but for which the decision system (31) still has a solution.

Hence, we also obtain the following existence theorem for closed-loop decision rules obtained through sharp convex selection procedures.

THEOREM 5.4. *Let S_G be a convex selection of the set-valued map G. Then, for any initial state* $(z_0, u_0, v_0) \in \text{graph}\,(R)$, *there exists a starting at* (z_0, u_0, v_0) *to the associated system of differential inclusions*

$$(32) \qquad \begin{aligned} &\text{(i)} \quad z'(t) = h(z(t), u(t), v(t)), \\ &\text{(ii)} \quad (u'(t), v'(t)) \in S(DR(z(t), u(t), v(t))h(z(t), u(t), v(t))) \\ &\qquad\qquad := G(z(t), u(t), v(t)) \cap S_G(z(t), u(t), v(t)). \end{aligned}$$

In particular, if we assume further that the selection procedure S_G is single-valued, then the single-valued map

$$(\tilde{c}(z, u, v), \tilde{d}(z, u, v)) := S(G)(z, u, v)$$

is a closed-loop decision rule, for which decision system (31) *has a solution for any initial state* $(z_0, u_0, v_0) \in \text{graph}\,(R)$.

Proof. We shall replace the system of differential inclusions (29) by the system of differential inclusions

$$(33) \qquad \begin{aligned} &\text{(i)} \quad z'(t) = h(z(t), u(t), v(t)), \\ &\text{(ii)} \quad (u'(t), v'(t)) \in S_G(z(t), u(t), v(t)). \end{aligned}$$

Since the convex selection procedure S_G has a closed graph and convex values, the right-hand side is an upper semicontinuous set-valued map with nonempty compact convex images and with linear growth. It remains to check that Graph R is still a viability domain for this new system of differential inclusions. Indeed, by construction, we know that there exists an element w in the intersection of $G(z, u, v)$ and $S_G(z, u, v)$.

This means that the pair $(h(z, u, v), w)$ belongs to $h(z, u, v) \times S_G(z, u, v)$ and that it also belongs to

$$\text{Graph}\,(G) := T_{\text{Graph}\,R}(z, u).$$

Therefore, we can apply Haddad's viability theorem. For any initial situation (z_0, u_0, v_0), there exists a solution $(z(\cdot), u(\cdot), v(\cdot))$ to the new system of differential inclusions (33) which is viable in Graph (R). Consequently, for almost all $t > 0$, the pair $(z'(t), u'(t), v'(t))$ belongs to the contingent cone to the graph of R at $(z(t), u(t), v(t))$, which is the graph of the contingent derivative $DR(z(t), u(t), v(t))$. In other words,

$$\text{for almost all } t > 0, \quad (u'(t), v'(t)) \in G(z(t), u(t), v(t)).$$

We thus deduce that for almost all $t > 0$, $(u'(t), v'(t))$ belongs to the selection $S(G)(z(t), u(t), v(t))$ of the set-valued map $G(z(t), u(t), v(t))$. Hence, we have found a solution to the system of differential inclusions (32). □

We can now multiply the possible corollaries, since we have given several instances of selection procedures of set-valued maps.

Example. Cooperative behavior. Let $\sigma : \text{Graph}\,(G) \mapsto \mathbf{G}$ be continuous.

COROLLARY 5.1. *Let us assume that the set-valued map G is lower semicontinuous with nonempty closed convex images on* Graph (R). *Let σ be continuous on* Graph (G) *and convex with respect to the pair (u, v). Then, for all initial situations $(u_0, v_0) \in R(z_0)$, there exists a solution starting at (z_0, u_0, v_0) and to the differential game* (3)-(21) *which is regulated by*:

For almost all $\geqq 0$, $(u'(t), v'(t)) \in G(z(t), u(t), v(t))$ *and* $\sigma(z(t), u(t), v(t), u'(t), v'(t)) = \inf_{u', v' \in G(z(t), u(t), v(t))} \sigma(z(t), u(t), v(t), u', v')$.

In particular, the game can be played by the heavy decision of minimal norm:

$$(c^0(z, u, v), d^0(z, u, v)) \in G(z, u, v),$$
$$\|c^0(z, u, v)\|^2 + \|d^0(z, u, v)\|^2) = \min_{(u', v') \in G(z, u, v)} (\|u'\|^2 + \|v'\|^2).$$

Example. Noncooperative behavior. We can also choose strategies in the regulation sets $G(z, u, v)$ in a noncooperative way, as saddle points of a function $a(z, u, v, \cdot, \cdot)$.

COROLLARY 5.2. *Let us assume that the set-valued map G is lower semicontinuous with nonempty closed convex images on* Graph (R) *and that a $\mathbf{R}^n \times \mathbf{R}^p \times \mathbf{R}^q \to \mathbf{R}$ satisfies that*

(i) *a is continuous,*

(ii) *For all (z, u, v, d), $c \mapsto a(z, u, v, c, d)$ is convex,*

(iii) *For all (z, u, v, c), $d \mapsto a(z, u, v, c, d)$ is concave.*

Then, for all initial situations $(u_0, v_0) \in R(z_0)$, there exists a solution starting at (z_0, u_0, v_0) and to the differential game (3)-(21) *which is regulated by*

$$\text{for almost all } t \geqq 0, \begin{cases} \text{(i)} & (u'(t), v'(t)) \in G(z(t), u(t), v(t)), \\ \text{(ii)} & \text{For all } (u', v') \in G(z(t), u(t), v(t)), \\ & a(z(t), u(t), v(t), u'(t), v') \\ & \qquad \leqq a(z(t), u(t), v(t), u'(t), v'(t)) \\ & \qquad \leqq a(z(t), u(t), v(t), u', v'(t)). \end{cases}$$

Appendix. Lower semicontinuous Lyapunov functions. We now consider a differential inclusion

$$\text{for almost all } t \geqq 0, \quad x'(t) \in F(x(t)) \tag{34}$$

and time-dependent functions $w(\cdot)$ defined as solutions to a differential equation

$$w'(t) = -\varphi(w(t)), \qquad w(0) = V(x(0)) \tag{35}$$

where $\varphi : \mathbf{R}_+ \to \mathbf{R}$ is a given continuous function with linear growth. This function φ is used as a parameter in what follows.

The main instance of such a function φ is the affine function $\varphi(w) := aw - b$, the solutions of which are $w(t) = (w(0) - b/a)\, e^{-at} + b/a$.

Our problem is to characterize either φ-Lyapunov functions, i.e., nonnegative extended[8] functions $V : X \to \mathbf{R}_+ \cup \{+\infty\}$ satisfying

$$\forall t \geqq 0, \quad V(x(t)) \leqq w(t), \quad w(0) = V(x(0)) \tag{36}$$

along at least a solution to the differential inclusion (34) or φ-universal Lyapunov functions, which satisfy property (36) along all solutions to (34).

DEFINITION 6.1. We shall say that a nonnegative contingently epidifferentiable[9] extended function V is a *Lyapunov function* of F associated with a function $\varphi(\cdot) : \mathbf{R}_+ \mapsto \mathbf{R}$ if and only if V is a solution to the contingent Hamilton-Jacobi inequalities

$$\forall x \in \mathrm{Dom}\,(V), \quad \inf_{v \in F(x)} D_\uparrow V(x)(v) + \varphi(V(x)) \leqq 0 \tag{37}$$

and a *universal Lyapunov function* of F associated with a function φ if and only if V is a solution to the upper contingent Hamilton-Jacobi inequalities

$$\forall x \in \mathrm{Dom}\,(V), \quad \sup_{v \in F(x)} D_\uparrow V(x)(v) + \varphi(V(x)) \leqq 0. \tag{38}$$

THEOREM 6.1. *Let V be a nonnegative contingently epidifferentiable extended function and $F : X \rightsquigarrow X$ be a nontrivial set-valued map.*

— Let us assume that F is upper semicontinuous with compact convex images and linear growth. Then V is a Lyapunov function of F associated with $\varphi(\cdot)$ if and only if for all initial state $x_0 \in \mathrm{Dom}\,(V)$, there exist solutions $x(\cdot)$ to differential inclusion (34) *and $w(\cdot)$ to differential equation* (35) *satisfying property* (36).

— If F is Lipschitz on the interior of its domain with compact values, then V is a universal Lyapunov function associated with φ if and only if for all initial state $x_0 \in \mathrm{Dom}\,(V)$, all solutions $x(\cdot)$ to differential inclusion (34) *and $w(\cdot)$ to differential equation* (35) *do satisfy property* (36).

The proof is based on the viability and invariance theorems of the closed subset Ep V for the differential inclusion:

$$\begin{aligned} &\text{(i)} \quad x'(t) \in F(x(t)), \\ &\text{(ii)} \quad w'(t) = -\varphi(w(t)) \end{aligned} \tag{39}$$

and these viability and invariance theorems can be reformulated in the following way.

COROLLARY 6.1. *Let $F : X \rightsquigarrow X$ be a nontrivial set-valued map.*

— Let us assume that F is upper semicontinuous with compact convex images and linear growth.

---

[8] In Rockafellar's sense.

[9] This means that for all $x \in \mathrm{Dom}\,(V)$ and $v \in X$, $D_\uparrow V(x)(v) > -\infty$, and that $D_\uparrow V(x)(v) < \infty$ for at least a $v \in X$.

A closed subset K enjoys the viability property if and only if its indicator Ψ_K is a solution to the contingent equation

$$\inf_{v \in F(x)} D_\uparrow \Psi_K(x)(v) = 0.$$

— If F is Lipschitz on the interior of its domain with compact values, then K is invariant by F if and only if its indicator Ψ_K is a solution to the contingent equation

$$\sup_{v \in F(x)} D_\uparrow \Psi_K(x)(v) = 0.$$

The functions φ and $U : X \to \mathbf{R}_+ \cup \{+\infty\}$ being given, can we construct the smallest lower semicontinuous Lyapunov function of a set-valued map F associated with φ larger than or equal to U, i.e., the smallest nonnegative lower semicontinuous solution U_φ to the contingent Hamilton-Jacobi inequalities (37) larger than or equal to U?

THEOREM 6.2. *Let us consider a nontrivial set-valued map $F : X \rightsquigarrow X$, a continuous function $\varphi : \mathbf{R}_+ \to \mathbf{R}$ with linear growth, and a proper nonnegative extended function U.*

— Let us assume that F is upper semicontinuous with compact convex images and linear growth. Then there exists a smallest nonnegative lower semicontinuous solution $U_\varphi : \mathrm{Dom}(F) \mapsto \mathbf{R} \cup \{+\infty\}$ to the contingent Hamilton-Jacobi inequalities (37) *larger than or equal to U (which can be the constant $+\infty$), which then enjoys the property*:

For all $x \in \mathrm{Dom}(U_\varphi)$, there exists solutions to (35) *and* (36) *satisfying for all $t \geqq 0$, $U(x(t)) \leqq U_\varphi(x(t)) \leqq w(t)$.*

— If F is Lipschitz on the interior of its domain with compact values and φ is Lipschitz, then there exists a smallest nonnegative lower semicontinuous solution $\tilde{U}_\varphi : \mathrm{Dom}(F) \mapsto \mathbf{R} \cup \{+\infty\}$ to the upper contingent Hamilton-Jacobi inequalities (35) *larger than or equal to U (which can be the constant $+\infty$), which then enjoys the property*:

For all $x \in \mathrm{Dom}(U_\varphi)$, all solutions to (35) *and* (36) *satisfy for all $t \geqq 0$, $U(x(t)) \leqq U_\varphi(x(t)) \leqq w(t)$.*

In particular, for $\varphi(w) := aw$, we deduce that

For all $x \in \mathrm{Dom}(U_a)$, $U(x(t)) \leqq U_a(x_0)\, e^{-at}$ and thus converges to zero.

The proof amounts to showing that the largest closed viability domain (invariance domain) contained in the epigraph of U, called the viability kernel (invariance kernel) of $\mathrm{Ep}(U)$, which does exist under the assumptions of the first (second) part of the theorem, is actually an epigraph, and thus, the one of the smallest lower semicontinuous (universal) Lyapunov function. Actually, the existence theorems of these kernels are equivalent to the theorem above, since it implies the following corollary.

COROLLARY 6.2. *We posit the assumptions of Theorem* 6.2.

— Let us assume that F is upper semicontinuous with compact convex images and linear growth.

The indicator $\Psi_{\mathrm{Viab}(K)}$ of the viability kernel $\mathrm{Viab}(K)$ *of a closed subset K (i.e., the largest closed viability domain of F contained in K) is the smallest nonnegative lower semicontinuous solution to*

$$\forall x \in \mathrm{Dom}(V), \quad \inf_{v \in F(x)} D_\uparrow V(x)(v) \leqq 0 \tag{40}$$

larger than or equal to Ψ_K.

— *Assume that F is Lipschitz on the interior of its domain with compact values.*

The indicator $\Psi_{\mathrm{Inv}(K)}$ of the invariant kernel $\mathrm{Inv}(K)$ *of a closed subset K (i.e., the largest closed invariance domain of F contained in K) is the smallest nonnegative lower semicontinuous solution to*

$$\forall x \in \mathrm{Dom}(V), \quad \sup_{v \in F(x)} D_{\uparrow} V(x)(v) \leqq 0 \tag{41}$$

larger than or equal to Ψ_K.

REFERENCES

[1] J.-P. AUBIN, *Viability tubes and the target problem*, in Modelling and Adaptive Control, C. Byrnes and A. Kurzhanski, eds., Lecture Notes in Control and Information Sci. 105, Springer-Verlag, New York, Berlin, 1988.

[2] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Grundlehren der Math. Wissenschaften, Vol. 264, Springer-Verlag, Berlin, New York, 1984, pp. 1-342.

[3] J.-P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.

[4] J.-P. AUBIN AND H. FRANKOWSKA, *Heavy viable trajectories of controlled systems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 2 (1985), pp. 371-395.

[5] ———, *Observability of systems under uncertainty*, SIAM J. Control Optim., 27 (1989), pp. 949-975.

[6] ———, *Set-Valued Analysis*, Birkäuser Verlag, Basel, 1990.

[7] J.-P. AUBIN AND R. WETS, *Stable approximations of set-valued maps*, Ann. Inst. H. Poincaré Anal. Nonlinéaire, 5 (1988), pp. 519-535.

[8] L. D. BERKOWITZ, *A variational approach to differential games*, in Advances in Game Theory, Princeton University Press, Princeton, NJ, 1964, pp. 127-174.

[9] P. BERNHARD, *Contribution à l'étude des jeux différentiels à somme nulle et information parfaite*, Thèse, Université de Paris VI, Paris, France, 1979.

[10] ———, *Exact controllability of perturbed continuous-time linear systems*, Trans. Automat. Control, 25 (1980), pp. 89-96.

[11] ———, *Differential games: introduction*, in Systems and Control Encyclopedia, Theory, Technology Applications, M. G. Singh, ed., Pergamon Press, London, 1988.

[12] ———, *Differential games: Isaacs' equations*, in Systems and Control Encyclopedia, Theory, Technology Applications, M. G. Singh, ed., Pergamon Press, London, 1988.

[13] ———, *Differential games: closed loop*, in Systems and Control Encyclopedia, Theory, Technology Applications, M. G. Singh, ed., Pergamon Press, London, 1988.

[14] ———, *Differential games: open loop*, in Systems and Control Encyclopedia, Theory, Technology Applications, M. G. Singh, ed., Pergamon Press, London, 1988.

[15] ———, *Differential games: linear quadratic*, in Systems and Control Encyclopedia, Theory, Technology Applications, M. G. Singh, ed., Pergamon Press, London, 1988.

[16] A. BLAQUIERE, F. GERARD, AND G. LEITMANN, *Quantitative and Qualitative Games*, Academic Press, New York, 1969.

[17] A. BLAQUIERE, *Dynamic games with coalitions and diplomacies*, in Directions in Large-Scale Systems, 1976, pp. 95-115.

[18] J. V. BREAKWELL, *Zero-sum differential games with terminal payoff*, in Differential Games and Applications, P. Hagedorn, H. W. Knobloch, and G. H. Olsder, eds., Lecture Notes in Control and Information Sciences, Vol. 3, Springer-Verlag, Berlin, New York, 1977.

[19] M. CORLESS AND G. LEITMANN, *Adaptive control for uncertain dynamical systems*, in Dynamical Systems and Microphysics Control Theory and Mechanics, 1984, pp. 91-158.

[20] M. CORLESS, G. LEITMANN, AND E. P. RYAN, *Tracking in the presence of bounded uncertainties*, in Proc. 4th International Conference on Control Theory, 1984.

[21] M. FALCONE AND P. SAINT-PIERRE, *Slow and quasi-slow solutions of differential inclusions*, J. Nonlinear Anal. Theory, Methods, Applications, 3 (1987), pp. 367-377.

[22] W. FLEMING, *The convergence problem for differential games*, J. Math. Anal. Appl., 3 (1961), pp. 102-116.

[23] H. FRANKOWSKA, *L'équation d'Hamilton-Jacobi contingente*, Comptes Rendus de l'Académie des Sciences, Paris, France, 1987.

[24] H. FRANKOWSKA, *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi*, Appl. Math. Optim., 19 (1989), pp. 291-311.

[25] H. G. GUSEINOV, A. I. SUBBOTIN, AND V. N. USHAKOV, *Derivatives for multivalued mappings with applications to game theoretical problems of control*, Problems Control Inform. Theory, 14 (1985), pp. 155-167.

[26] G. HADDAD, *Monotone trajectories of differential inclusions with memory*, Israel J. Math., 39 (1981), pp. 38-100.

[27] O. HAJEK, *Pursuit Games*, Academic Press, New York, 1975.

[28] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[29] N. N. KRASOVSKI, *The Control of a Dynamic System*, Nauka, Moscow, 1986.

[30] N. N. KRASOVSKI AND A. I. SUBBOTIN, *Positional Differential Games*, Nauka, Moscow, 1974.

[31] A. B. KURZHANSKII, *Control and Observation Under Conditions of Uncertainty*, Nauka, Moscow, 1974.

[32] ———, *On the analytical properties of viability tubes of trajectories of differential systems*, Dokl. Acad. Nauk SSSR, 287 (1986), pp. 1047-1050.

[33] A. B. KURZHANSKII AND T. F. FILIPPOVA, *On the description of the set of viable trajectories of a differential inclusion*, Soviet Math. Dokl., 34 (1987), pp. 30-33.

[34] YU. S. LEDYAEV, *Regular differential games with mixed constraints on the controls*, in Proc. Steklov Institute of Mathematics, 167 (1985), pp. 233-242.

[35] G. LEITMANN, *Guaranteed avoidance strategies*, J. Optim. Theory Appl., 32 (1980), pp. 569-576.

[36] ———, *The Calculus of Variations and Optimal Control*, Plenum Press, New York, 1981.

[37] G. LEITMANN, E. P. RYAN, AND A. STEINBERG, *Feedback control of uncertain systems: robustness with respect to neglected actuator and sensor dynamics*, Internat. J. Control, 43 (1986), pp. 1243-1256.

[38] P.-L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.

[39] E. MICHAEL, *Continuous selections* I, Ann. of Math., 63 (1956), pp. 361-381.

[40] ———, *Continuous selections* II, Ann. of Math., 64 (1956), pp. 562-580.

[41] ———, *Continuous selections* III, Ann. of Math., 65 (1957), pp. 375-390.

[42] M. QUINCAMPOIX, *Playable differentiable games*, J. Math. Anal. Appl., to appear.

[43] A. RAY AND A. BLAQUIERE, *Sufficient conditions for optimality of threat strategies in a differential game*, J. Optim. Theory Appl., 33 (1988), pp. 99-109.

[44] A. I. SUBBOTIN, *Conditions for optimality of a guaranteed outcome in game problems of control*, in Proc. Steklov Institute of Mathematics, 167 (1985), pp. 291-304.

[45] A. I. SUBBOTIN AND N. N. SUBBOTINA, *Differentiability properties of the value function of a differential game with integral terminal costs*, Problems Control Inform. Theory, 12 (1983), pp. 153-166.

[46] A. I. SUBBOTIN AND A. M. TARASYEV, *Stability properties of the value function of a differential game and viscosity solutions of Hamilton-Jacobi equations*, Problems Control Inform. Theory, 15 (1986), pp. 451-463.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1321-1349, November 1990

004

ASYMPTOTIC STABILIZATION OF A CLASS OF SMOOTH TWO-DIMENSIONAL SYSTEMS*

W. P. DAYAWANSA†, C. F. MARTIN‡, AND G. KNOWLES§

Abstract. This paper studies the asymptotic stabilizability of two-dimensional control systems. The class under consideration includes C^∞-systems that satisfy a certain genericity assumption and all real analytic systems. Necessary and sufficient conditions for feedback stabilization using continuous feedback and a sufficient condition for C^1-feedback stabilization are given. This latter condition is given in terms of an inequality involving two indices. If the direction of the inequality is changed, an obstruction to C^∞-feedback stabilizability is obtained. A subclass of polynomial systems is also studied and given complete necessary and sufficient conditions for global asymptotic stabilization using C^1-feedback.

Key words. asymptotic stabilization, nonlinear systems, two-dimensional, Weierstrass polynomial systems

AMS(MOS) subject classification. 93

1. Introduction.

Asymptotic stabilization of a nonlinear control system is one of the most important problems in control theory. Fortunately, techniques have been developed in the recent past to analyze this problem. Prominent among them are the techniques based on center manifold theory, pioneered by Ayels [4] and used effectively by Kokotovic and his coauthors, among others; the idea of zero dynamics introduced by Byrnes and Isidori [9], [8], [7], etc.; the topological obstructions derived by Brockett [6], Krosnosel'skii and Zabreiko [20]; and the work on continuous feedback stabilization by Sontag and Sussmann [25], Kawski [18], etc.

In this paper attention is restricted to a smooth two-dimensional system,

$$\dot{x} = \tilde{f}(x) + \tilde{g}(x)u, \tag{1.1}$$

where $x \in U$ is an open subset of $\mathbb{R}^2$, u is a scalar input, and $\tilde{f}$, $\tilde{g}$ are smooth vector fields. It is assumed that $\tilde{f}(0) = 0$, $\tilde{g}(0) \neq 0$. We study the existence of a feedback function α on U such that $x(0) = 0$ and the closed-loop system

$$\dot{x} = (\tilde{f} + \tilde{g}\alpha)(x) \tag{1.2}$$

is asymptotically stable at zero in the sense of Lyapunov. Let us denote by $t \mapsto x(t, x^0)$ the solution of (1.2) with initial condition x^0. Here we recall that (1.2) is said to be stable at zero if for all $\varepsilon > 0$ there exist $\delta > 0$ such that $\|x(t, x^0)\| < \varepsilon$ for all $t > 0$ whenever $\|x^0\| < \delta$. The system (1.2) is said to be asymptotically stable at zero if it is stable and $x(t, x^0)$ converges to the origin as $t \to \infty$, for all x^0 in some neighborhood of the origin.

The recent work of Kawski [18] has shown that if the system is small time locally controllable at the origin, then it is stabilizable by Hölder continuous feedback. He constructed a class of Lyapunov functions to prove the asymptotic stability of the closed-loop system.

* Received by the editors December 27, 1988; accepted for publication (in revised form) November 29, 1989.

† Department of Mathematics, Texas Tech University, Lubbock, Texas 79409. The research of this author was supported in part by National Science Foundation grant ECS-8802483. Present address, Department of Electrical Engineering, Systems Research Center, University of Maryland, College Park, Maryland 20742.

‡ Department of Mathematics, Texas Tech University, Lubbock, Texas 79409. The research of this author was supported in part by National Security Agency grant MDA904-85-H0009.

§ Grumman Corporate Research Center, Grumman Corporation, Bethpage, New York, New York 11714-3580.

An extremely important observation on asymptotic stabilization was made by Brockett [6]. For the moment let us consider (1.1) with arbitrary state-space dimension n and arbitrary number of inputs n. Brockett proved that the following are necessary for stabilization of (1.1) with a C^1 feedback function.

(B1) The uncontrollable eigenvalues of the linearized system should be in the closed left half of the complex plane.

(B2) (1.1) is locally asymptotically controllable to the origin, i.e., for an arbitrary open neighborhood W of the origin there exist a neighborhood W of the origin and control $u(\cdot)$ such that for all $x^0 \in W$ the solution $t \mapsto x(t, x^0, u(t))$ of (1.1) stays in U for all $t > 0$ and converges to the origin as $t \mapsto \infty$.

(B3) The function $(x, u) \mapsto \tilde{f}(x) + \tilde{g}(x)u : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is locally onto at $(0, 0)$.

In a recent paper by Boothby and Marino [5] it was pointed out that in the two-dimensional analytic case, (B3) can be stated in an apparently stronger form, thereby deriving a new necessary condition.

The focus of this paper is on the two-dimensional case. We study the real analytic case and a class of smooth systems which are characterized by the property that a certain order jet approximates the system in a sense which will be described in § 2.

The main results of this paper are the following.

(i) System (1.1) (with $\tilde{f}(0) = 0$ and $\tilde{g}(0) \neq 0$ both $\tilde{f}$ and $\tilde{g}$ are C^w) is locally asymptotically stabilizable by continuous feedback if and only if (B2) is satisfied, i.e., the system is locally asymptotically controllable to the origin (Theorem 3.1, Corollary 3.1).

(ii) We will define two rational indices associated to (1.1), which will be called index of stabilizability and the fundamental stabilizability degree. We will show that if a certain inequality in terms of these two indices is satisfied, then the system is C^1 stabilizable. Furthermore, if they satisfy another inequality then the system is not C^∞ stabilizable (Theorems 4.1–4.3 and Corollary 4.1).

(iii) We consider the two-dimensional homogeneous case and give necessary and sufficient conditions for global asymptotic stabilizability (Theorem 5.2).

(iv) We consider the special class,

$$\dot{x}_1 = ax_1^n + bx_2^m, \qquad \dot{x}_2 = u,$$

when a, b are real numbers, n, m are positive integers, and we consider the global asymptotic stabilization problems (Theorem 6.1).

The basic mathematical ideas in the paper are as follows. First an equivalence relation on the space of all smooth two-dimensional systems is introduced. All elements of a given class will have the same stabilizability properties. Then it is possible to observe that if f is real analytic, then f is equivalent to a Weierstrass polynomial. Even when f is only C^∞, Mather's theory on singularities is applicable here and gives fairly weak sufficient conditions under which f is equivalent to a polynomial system. It turns out that if the linearization around the origin is nontrivial, then the system is equivalent to its linearization and in this sense the linearization technique is generalized.

Before embarking on the question of stabilizability of polynomial or Weierstrass polynomial systems, a special class of polynomial systems should be considered. This case will illuminate the picture considerably. The algebraic structure of a one-dimensional real analytic variety is used to give sufficient conditions for asymptotic stabilization by C^1 feedback and necessary and sufficient conditions for asymptotic stabilization by continuous feedback (which preserves existence and uniqueness of

solutions and has slower than exponential decay rate). The sufficient conditions mentioned above are shown to be the union of a necessary condition and an additional condition. Surprisingly, the additional condition is satisfied rather easily when the complexity of the system increases. An obstruction to stabilizability with C^∞-feedback is also given.

This paper is organized as follows. In § 2, an equivalence relation (weak feedback equivalence) with the property that equivalent systems have the same stabilizability properties is defined. This relation turns out to be the same as the one used by Golubitsky and Schäffer [13] in the context of bifurcation theory. Their results are used to identify the equivalence classes which contain polynomial or Weierstrass polynomial systems.

In § 3, the local asymptotic stabilization problem for real analytic systems,

$$\dot{x} = f(x_1, x_2), \qquad \dot{x}_2 = u,$$

is solved completely. The feedback used is continuous at the origin and C^∞—away from the origin. As has been pointed out by many authors ([25], [28], [15], etc.), this allows us to use C^∞-feedback in practical situations. In view of a theorem due to Artstein [3] (see [24] also), existence of continuous stabilizing feedback implies the existence of stabilizing feedback of the class considered above. However we use this stronger class of feedback right away due to the fact that it answers existence and uniqueness questions trivially. In § 4, some sharp sufficient conditions for C^1-stabilizability and some obstructions to C^∞-stabilizability are given. In § 5 we study the global stabilization problem for homogeneous systems. In § 6 we consider the class

$$\dot{x}_1 = ax_1^n + bx_2^m, \qquad \dot{x}_2 = u,$$

where $a, b \in \mathbb{R}$, and n, m are nonnegative integers. We give necessary and sufficient conditions for global stabilization by C^1 (and in many cases real analytic) feedback. This class already contains very interesting systems. For example, $\dot{x}_1 = x_1 - x_2^3$; $\dot{x}_2 = u$ is stabilizable by Hölder continuous feedback of Hölder exponent $\frac{1}{3}$ but no more (in particular not by Lipschitz continuous feedback) (see [12]). Also the system $\dot{x}_1 = x_1^2 - x_2^4$, $\dot{x}_2 = u$ is C^1-stabilizable but not C^3-stabilizable (see [12]).

In § 7 some concluding remarks are given.

2. Weak feedback equivalence and the orbits of polynomial and Weierstrass polynomial systems. In this section we identify a class of smooth systems which can be approximated by a Taylor polynomial of an appropriate order. The results of most of this section are not needed to analyze the real analytic class and the reader can skip to the last paragraph of the section without losing the essential flavor of this paper.

The usual notion of feedback equivalence [26] is defined as follows.

DEFINITION 2.1. Consider two two-dimensional systems $\dot{x} = f_i(x) + g_i(x)u_i$ where $i = 1, 2$, $x \in U$, an open neighborhood of the origin in $\mathbb{R}^2$, f_i and g_i are smooth vector fields, and $f_i(0) = 0$. The two systems are feedback equivalent near the origin if there exist germs of real-valued smooth functions α and β near the origin, where $\alpha(0) = 0$, $\beta(0) \neq 0$ and a germ φ of a diffeomorphism around the origin which preserves the origin such that

$$f_1 = \varphi_*(f_2 + \alpha g_2), \tag{2.1}$$

$$g_1 = \varphi_*(\beta g_2). \tag{2.2}$$

This notion is of tremendous help in understanding controllability properties since in many instances we can simplify the structure using this equivalence relation. Obviously, stabilizability properties of feedback equivalent systems are the same.

Indeed, we can enlarge the equivalence relation without losing this property by allowing the multiplication of the vector fields f_i by germs of positive smooth functions at the origin.

Here the goal is to preserve the structure (2.3) of the system and thus the diffeomorphism is restricted to a subgroup.

DEFINITION 2.2. Consider two systems,

$$\dot{x}_1 = f^i(x), \quad \dot{x}_2 = u, \quad i = 1, 2, \tag{2.3}$$

where f^i is a smooth function in a neighborhood of the origin. The two systems (or by abuse of language f^1 and f^2) are weakly feedback equivalent, if there exists a germ of a coordinate transformation around the origin of the form $(x_1, x_2) \mapsto (\varphi(x_1), \psi(x_1, x_2))$, which preserves the origin, and a function γ such that

$$\gamma(0) \frac{d\varphi}{dx_1}(0) > 0,$$

and

$$f^1(x_1, x_2) = \gamma(x) f^2(\varphi(x_1), \psi(x_1, x_2)), \tag{2.4}$$

as germs of smooth functions at the origin.

This equivalence relation is up to sign, the same as that defined in Golubitsky and Schäffer [13, p. 51] in the context of bifurcation theory. Following the work of Thom [27] and Mather [23], they raised the question of finding conditions under which a given function is equivalent (weak feedback equivalent in this paper) to a Taylor polynomial. The relevant results are described below.

Let $\mathcal{G}$ denote the space of germs of real-valued smooth functions in two variables at the origin. $\mathcal{M}$ will denote the subset of $\mathcal{G}$ consisting of germs which take the value zero at the origin. $\mathcal{M}$ is clearly an ideal in $\mathcal{G}$. Henceforth, the term germ refers to an element of $\mathcal{G}$.

DEFINITION 2.3 [13]. The restricted tangent space of a germ g, denoted by $RT(g)$, is the ideal in $\mathcal{G}$ generated by $\{g, x_1(\partial g/\partial x_2), x_2(\partial g/\partial x_2)\}$.

THEOREM 2.1 [13]. *Let $g, p \in \mathcal{G}$. If $RT(g + tp) = RT(g)$ for all $t \in [0, 1]$, then $g + tp$ is weakly feedback equivalent to g for all $t \in [0, 1]$.*

THEOREM 2.2 [13]. *Let $g \in \mathcal{M}$.*

(i) *There exists an integer k such that $\mathcal{M}^k \subset RT(g)$ if and only if* $\dim(\mathcal{M}/RT(g))$ *is finite.*

(ii) *If $\mathcal{M}^k \subset RT(g)$ for some integer k, then g is weakly feedback equivalent to its kth order Taylor polynomial.*

Theorem 2.2 can be used to identify a large number of equivalence classes which contain polynomial systems. For more general means of identifying these classes and for means of computing the associated transformations, we refer the reader to Golubitsky and Schäffer [13].

In order to give the reader some idea of the results given in later sections an example, the case $\dim(RT(f)) \leqq 3$, is considered. Since the linearization techniques apply when the origin is a regular point of f, it is assumed that it is a critical point. The computation of the normal forms and the means of identification of the normal form was done in Golubitsky and Schäffer [13]. Results of § 4 are used to determine the asymptotic stabilizability properties. We remark here that cases (1), (4), (7), and (8) follow from the results of Boothby and Marino in [5], and the cases (2) and (3) are obvious. The purpose of Table 2.1 is merely to motivate the reader rather than to present new results here.

TABLE 2.1

The possible cases of f when $\dim(RT(f)) \leqq 3$, *normal forms, and the asymptotic stabilizability with* C^1*-feedback.*

	Defining relation	Normal form	Asymptotic stabilizability
(1)	$\det(d^2f(0)) < 0$; $f_{x_2x_2}(0) \neq 0$	$\pm(x_2^2 - x_1^2)$	Asymptotically stabilizable
(2)	$\det(d^2f(0)) > 0$; $f_{x_2x_2}(0) \neq 0$	$\pm(x_1^2 + x_2^2)$	Not asymptotically stabilizable
(3)	$\det(d^2f(0)) = 0$, $\exists v \neq 0$ such that $f_{vv}(0) = 0$ and $f_{vvv}(0) \neq 0$; $f_{x_2x_2}(0) \neq 0$	$x_2^2 \pm x_1^3$	$x_2^2 - x_1^3$ is asymptotically stabilizable $x_2^2 + x_1^3$ is not asymptotically stabilizable
(4)	$f_{x_2x_2}(0) = 0$, $f_{x_2x_2x_2}(0) \neq 0$, $f_{x_1x_2}(0) \neq 0$	$\pm x_2^3 \pm x_1x_2$	Asymptotically stabilizable
(5)	$\det(d^2f(0)) = 0$; $\exists v \neq 0$ such that $f_{vv}(0) = f_{vvv}(0) = 0$; $p = f_{x_2x_2}(0) \neq 0$, $q = f_{vvvv}(0)f_{x_2x_2}(0) - 3f_{vvx_2}^2(0) \neq 0$, p and q have same signs.	$\pm(x_1^4 + x_2^2)$	Not asymptotically stabilizable
(6)	Same as in (5) except that p and q have opposite signs	$\pm(x_1^4 - x_2^2)$	Asymptotically stabilizable
(7)	$f_{x_2}(0) = f_{x_1x_2}(0) = f_{x_1x_1}(0) \neq 0 \neq f_{x_2x_2x_2}(0)$	$\pm x_1^2 \pm x_2^3$	Asymptotically stabilizable
(8)	$f_{x_2x_2x_2}(0) = 0$, $f_{x_2x_2x_2x_2}(0) \neq 0$, $f_{x_1x_2}(0) \neq 0$	$\pm x_1x_2 \pm x_2^4$	Asymptotically stabilizable

If f is real analytic, then f is always weakly feedback equivalent to a Weierstrass polynomial up to sign. To see this, first use weak feedback equivalence with $\gamma(x) = 1$, $\varphi(x_1) = x_1$, and $\psi(x_1, x_2) = ax_1 + bx_2$ (for some $a, b \in \mathbb{R}$) to ensure that $f(x_1, 0) \not\equiv 0$. Now we consider complexity f. For the sake of clarity, use z_1 and z_2 for the complex variables instead of x_1 and x_2. Now it is well known (see, e.g., Griffiths and Harris [14]) that there is a unique holomorphic function $\mathcal{G}(x_1, z_2)$ such that $\mathcal{G}(0) \neq 0$ and a unique Weierstrass polynomial $z_1^m + a_1(z_2)z_1^{m-1} + \cdots + a_m(z_2)$ (where $a_i(\cdot)$ is holomorphic and $a_i(0) = 0$, $i = 1, \cdots, m$) such that $f(z_1, z_2) = \mathcal{G}(z_1, z_2)(z_1^m + a_1(z_2)z_1^{m-1} + \cdots + a_m(z_2))$. It is now possible to claim that $\mathcal{G}(x_1, x_2)$ and $a_i(x_2)$, $i = 1, \cdots, m$ are all real. For if not, then (since $f(x_1, x_2)$ is real)

$$f(z_1, z_2) = \tilde{\mathcal{G}}(z_1, z_2)(z_1^m + \tilde{a}(z_2)z_1^{m-1} + \cdots + \tilde{a}_m(z_2)),$$

where $\tilde{\mathcal{G}}(z_1, z_2)$ and $\tilde{a}_i(z_2)$ denotes the complexification of the complex conjugates of $\mathcal{G}(x_1, x_2)$ and $a_i(x_2)$. But this violates the uniqueness of $\mathcal{G}$ and a_i.

Now $f(x_1x_2) = \mathcal{G}(x_1, x_2)(x_1^m + a_1(x_2)x_1^{m-1} + \cdots + a_m(x_2))$ and $\mathcal{G}(0) \neq 0$. Therefore, f is weakly feedback equivalent to $\operatorname{sgn}(\mathcal{G}(0))(x_1^m + a_1(x_2)a_1^{m-1} + \cdots + a_m(x_2))$ as previously claimed.

3. Asymptotic stabilization of two-dimensional systems. In this section we will give necessary and sufficient conditions for continuous local feedback stabilization of the real analytic system

$$\dot{x} = f(x_1, x_2), \qquad \dot{x}_2 = u, \tag{3.1}$$

where $f(0) = 0$.

THEOREM 3.1. *System* (3.1) *is asymptotically stabilizable with continuous feedback if and only if for all* $\varepsilon > 0$ *there exist* $p \in B_\varepsilon(0) \cap \mathbb{R}_+^2$ *and* $q \in B_\varepsilon(0) \cap \mathbb{R}_-^2$ *such that* $f(p) < 0$ *and* $f(q) > 0$. *This condition is also equivalent to local asymptotic controllability to the origin.*

COROLLARY 3.1. *System* (3.1) *with smooth* f *is asymptotically stabilizable if* f *is weakly feedback equivalent to a polynomial and if for all* $\varepsilon > 0$ *there exist* $p \in B_\varepsilon(0) \cap \mathbb{R}_+^2$ *and* $q \in B_\varepsilon(0) \cap \mathbb{R}_-^2$ *such that* $f(p) < 0$ *and* $f(q) > 0$.

The equivalence of the given condition in Theorem 3.1 to local asymptotic controllability is trivial in view of Facts 1-5 to follow. In the rest of the section we will prove Theorem 3.1. Corollary 3.1 follows at once from Theorem 3.1.

First consider a two-dimensional system,

$$\dot{x}_1 = \pm f(x_1, x_2), \qquad \dot{x}_2 = u, \tag{3.2}$$

where $f(x_1, x_2)$ is a Weierstrass polynomial, i.e.,

$$f(x_1, x_2) = x_1^n + a_1(x_2)x_1^{n-1} + \cdots + a_n(x_2), \quad \text{and}$$

$a_i(0) = 0$, $1 \leqq i \leqq n$. It was shown in § 2 that all two-dimensional real analytic systems and a large class of two-dimensional smooth systems contain a system of the form (3.2) in their orbits under the weak feedback equivalence. Since asymptotic stabilizability is invariant under the equivalence relation, necessary and sufficient conditions for asymptotic stabilizability of (3.2) will give necessary and sufficient conditions for the asymptotic stabilizability of all two-dimensional real analytic systems and for a large class of smooth two-dimensional systems which satisfy the hypothesis of Theorem 2.2.

Here only the case with the positive sign in (3.2) will be discussed. The remaining case can be treated in the same way. Moreover, since the local problem is considered, only the germs of the corresponding functions at the origin will be dealt with. So questions such as the convergence of a series, etc., should be interpreted in this context.

The basic steps of the proof are as follows: First, it will be observed that (possibly after redefining the x_1-axis) $f^{-1}(0)$ may be written as a finite union of graphs of C^1 functions $\{x_1 = \lambda_i(x_2)\}_{i \in I}$ where each λ_i has domain either $[0, \varepsilon)$ or $(-\varepsilon, 0]$. Moreover, each of these functions is either strictly monotone or identically zero; $\lambda_i(0) = 0$ and has a representation as a convergent rational power series in its argument. Therefore, in a small neighborhood of the origin these curves do not intersect each other except at the origin and, hence, describe sectors in this neighborhood. These sectors will be modified appropriately, linear feedback functions in these sectors will be defined, and then they will be pieced together using a smooth partition of unity and it will be shown that the resulting feedback will asymptotically stabilize the system. This is done by constructing a neighborhood base of the origin $\{W^\beta\}_{\beta < \beta_0}$ such that each W^β is positively invariant, contains no nontrivial periodic orbits, and no equilibrium points other than at the origin. Then, by invoking the Poincaré-Bendixon theory, it is possible to prove the asymptotic stability. Since the partition of unity is defined on a deleted neighborhood of the origin, feedback functions will be C^∞ only on a deleted neighborhood of the origin. However, since the magnitude of the feedback functions is less than $k\|x\|$, it extends to a continuous function in a neighborhood of the origin. This special structure of feedback obviously ensures the existence and uniqueness of solutions.

It should be observed that it follows from Artstein [3] (see [24] also) that when continuous stabilizing feedback exists, we can find stabilizing feedback functions from this smaller class. However, working with this class from the beginning avoids difficulties regarding existence and uniqueness of solutions.

First let x_1 and x_2 be considered complex variables and, for the sake of clarity, write them as z_1 and z_2. First consider the case when $f(z_1, z_2)$ is an irreducible polynomial $z_1^m + a_1(z_2)z_1^{m-1} + \cdots + a_m(z_2)$ where $a_i(0) = 0$ for all i. Now f defines an algebraic function $z_1 = \mathcal{G}(z_2)$ [1, p. 292] and we can write

$$f(z_1, z_2) = \prod_{i=1}^{m} (z_1 - \lambda_i(z_2)),$$

where $\lambda_i(z_2)$ are the branches of $\mathcal{G}(z_2)$. In particular, we may define λ_i to be holomorphic on $\{z \in \mathbb{C} \mid z \neq 0, \arg(z) \neq \pi/2\}$. Moreover, since $w \mapsto \lambda_i(w^m)$ is holomorphic on $\mathbb{C}$, it

follows that $\lambda_i(z)$ can be written as a convergent rational power series

$$\lambda_i(z) = \sum_{n=1}^{\infty} a_{i,n} z^{n/m}.$$

All of these considerations are valid when f is a Weierstrass polynomial except that the convergence is now local (see Lefschetz [21, p. 103]) (and the function defined by the analytic continuation of $\{\lambda_i\}_{i=1}^m$ is not an algebraic function).

We remark here that Boothby and Marino [5] used this algebraic scheme to obtain a necessary condition for stabilization.

The following salient features of $\{\lambda_i\}_{i=1}^m$ are noted. They are more or less obvious and the proofs are omitted. The statements made on $[0, \varepsilon)$ are valid on intervals $(-\varepsilon, 0]$, also.

FACT 1. For small $\varepsilon > 0$ and for $x_2 \in [0, \varepsilon]$, $\lambda_i(x_2)$ is either always real or it takes a real value only at $x_2 = 0$.

FACT 2. If $a_{i,n} = 0$ for all $n < m$, then the function $\lambda_i : (-\varepsilon, \varepsilon) \to \mathbb{C}$ is C^1 and $\lambda_i'(x_2) = \sum_{n=1}^{\infty} (n/m) a_{i,n} x_2^{((n/m)-1)}$.

FACT 3. Suppose that $\lambda_i | [0, \varepsilon)$ is real. (Now $a_{i,n}$ is real for all n.) If $a_{i,n} \neq 0$ for some $n < m$, then the equation $x_1 = \lambda_i(x_2)$ can be solved to obtain a convergent rational power series

$$x_2 = \sum_{n>l} b_n x_1^{n/l}, \quad x_2 \in [0, \delta) \quad \text{or} \quad x_2 \in (-\delta, 0].$$

The function μ_i defined by this function is C^1 and, hence, the graph of $x_1 = \lambda_i(x_2)(x_2 \in [0, \varepsilon))$ is a C^1 submanifold of $\mathbb{R}^2$. On the other hand, if $a_{i,n} = 0$ for all $n < m$, then $x_2 \mapsto \lambda_i(x_2)$ is C^1 on $[0, \varepsilon)$. Therefore, it follows that the curve $x_2 \mapsto (x_2, \lambda_i(x_2))$, $(x_2 \in [0, \varepsilon))$ is a C^1 submanifold. Therefore, in either case, tangent vector to this curve at zero is well defined. Now by redefining the x_1 axis, if necessary, it may be assumed that none of the tangent vectors to $x_1 = \lambda_i(x_2)$ at the origin is tangential to the x_1 axis. It now follows that if $\lambda_i | [0, \varepsilon)$ is real, then $a_{i,n} = 0$ for all $n < m$ and $x_2 \mapsto \lambda_i(x_2)$, $(x_2 \in [0, \varepsilon))$ is a C^1-curve. Henceforth, this condition, without loss of generality, will be assumed.

FACT 4. If $\lambda_i | [0, \varepsilon)$ is real, then $\lambda_i(x_2)$ is either strictly monotone for $x_2 \in [0, \varepsilon)$ or else $\lambda_i(x_2) \equiv 0$.

FACT 5. Let

$$\mathcal{C}_r = \{\text{graph of } x_1 = \lambda_j(x_2),\ x_2 \in [0, \varepsilon) \mid \lambda_j \text{ is a branch of } \mathcal{G} \text{ and } \lambda_j(x_2) > 0\ \forall x_2 \in (0, \varepsilon)\}$$

$$\cup \{\text{graph of } x_1 = \lambda_j(x_2),\ x_2 \in (-\varepsilon, 0] \mid \lambda_j \text{ is a branch of } \mathcal{G} \text{ and } \lambda_j(x_2) > 0\ \forall x_2 \in (-\varepsilon, 0)\}.$$

Let $\mathcal{C}_\ell$ be the corresponding collection of graphs obtained by replacing the condition $\lambda_j(x_2) > 0$ as above by $\lambda_j(x_2) < 0$. (The letters r and l stand for right and left.) An order can be defined on $\mathcal{C}_r$ (and on $\mathcal{C}_\ell$) by graph of $\lambda_i >$ graph of λ_j if $\lambda_i^{-1}(x_1) > \lambda_j^{-1}(x_1)$ for small enough x_1. Existence of λ_i^{-1} follows from the strict monotonicity and the independence of the definition on x_1 follows from the rational power series expansions.

In order to simplify the statements of the proofs to follow, the following notation and terminology will be adopted. Throughout, ε and δ will denote appropriately small positive reals and the graph of λ_i means graph of $x_1 = \lambda_i(x_2)$ for x_2 in a specified interval. $B_\varepsilon(0)$ will denote the Euclidean ball of radius $\varepsilon > 0$ around the origin in $\mathbb{R}^2$.

Let

$$\mathscr{C}_{r,u} = \{(\text{graph of } \lambda_i) \in \mathscr{C}_r \mid \text{dom}(\lambda_i) = [0, \varepsilon)\},$$

$$\mathscr{C}_{r,d} = \mathscr{C}_r - \mathscr{C}_{r,u},$$

$$\mathscr{C}_{l,u} = \{(\text{graph of } \lambda_i) \in \mathscr{C}_\ell \mid \text{dom}(\lambda_i) = [0, \varepsilon)\},$$

$$\mathscr{C}_{\ell,d} = \mathscr{C}_\ell - C_{\ell,u},$$

$$\mathscr{C} = \mathscr{C}_r \cup \mathscr{C}_\ell.$$

We will call $\sigma \in \mathscr{C}$ a *critical curve* if f changes sign while crossing σ. The set of critical curves will be denoted by $\mathscr{C}^c$ and $\mathscr{C}^c_{\alpha,\beta} = \mathscr{C}^c \cap \mathscr{C}_{\alpha,\beta}$ for $(\alpha, \beta) \in \{r, \ell\} \times \{u, d\}$. If $\sigma \in \mathscr{C}$ and it is not critical then σ is called a *noncritical curve.* Note that $f^{-1}(0) \cap B_\varepsilon(0) = \{x \in \sigma \mid \sigma \in \mathscr{C}\}$.

Let

$$\mathbb{R}^2_+ = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 > 0\},$$

$$\mathbb{R}^2_- = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 < 0\}.$$

Let $\sigma = \text{graph of } \lambda : [0, \varepsilon) \to [0, \infty)$ be in $\mathscr{C}^c_{r,u}$. Then it may be said that (x_1, x_2) is above σ if $0 < x_2 < \varepsilon$ and $x_1 < \lambda(x_2)$. The region above σ means the set of all x which is above σ. Terms such as "below σ" or "between σ_1 and σ_2" will be used in the sequel and they are self-explanatory.

The proof of the Theorem 3.1 will be broken into several different cases and in all of the interesting cases the feedback α will be defined sectorwise first. Then a smooth partition of unity on a deleted neighborhood of the origin will be used to piece them together. The asymptotic stability of the system will be proven by constructing a neighborhood base $\{W^\beta\}_{\beta \leqq \beta_0}$ of the origin such that for every point in each of these neighborhoods the positive limit set is the origin and that the neighborhoods are positively invariant. In order to construct these neighborhoods, the following preliminary facts are needed. The first lemma is needed later to show that the invariant sets $\{W^\beta\}_{\beta < \beta^0}$ form a neighborhood base of the origin.

LEMMA 3.1. *Let $\lambda : [0, \varepsilon) \to [0, \infty)$ be a C^1 function given by a convergent rational power series $\lambda(\theta) = \sum_{n=\ell}^{\infty} a_n \theta^{n/m}$ $(m, \ell \in \mathbb{N})$ and assume that $\lambda \neq 0$.*

Then for small enough β_0 there exists a function $\varphi : [0, \beta_0) \to [0, \infty)$ such that the following hold:

(i) *The graphs of $x_1 = \lambda(x_2)$ and of $x_2 = x_1 \ln(x_1/\beta)$ meet at $\varphi(\beta)$, $\varphi(\beta) \ln(\varphi(\beta)/\beta)$ and $\varphi(\beta)$ is the smallest such positive x_1 value.*

(ii) *$\varphi(\beta) \to 0$ as $\beta \to 0$.*

Proof. Without loss of generality it is assumed that $a_\ell \neq 0$. Since λ is C^1 at zero it follows that $\ell \geqq m$ and since λ is positive, $a_\ell > 0$. Now for small θ, $\frac{1}{2} a_\ell \theta^{\ell/m} \leqq \lambda(\theta) \leqq 2a_\ell \theta^{l/m}$. Since $x_1 \mapsto x_1 \ln(x_1/\beta)$ is monotone increasing on $x_1 \in [\beta, \infty)$ it suffices to prove that there are functions $\beta \mapsto \varphi_1(\beta)$ and $\beta \mapsto \varphi_2(\beta)$ $(\beta \in (0, \beta_0))$ such that $\varphi_1(\beta) \ln(\varphi_1(\beta)/\beta) = (1/2a_\ell)^{m/\ell} (\varphi_1(\beta))^{m/\ell}$;

$$\varphi_2(\beta) \ln\left(\frac{\varphi_2(\beta)}{\beta}\right) = \left(\frac{2}{a_\ell}\right)^{m/\ell} (\varphi_2(\beta))^{m/\ell} \quad \text{and} \quad \varphi_2(\beta) \to 0 \quad \text{as } \beta \to 0.$$

Now consider the equation

$$x_1 \ln\left(\frac{x_1}{\beta}\right) = b x_1^{m/\ell} \quad \text{where } b > 0 \text{ is fixed.}$$

This can be rearranged and written as

$$\beta(x_1) = x_1 \exp(-bx_1^{-(1-m/\ell)}).$$

Obviously the function $x_1 \mapsto \beta(x_1) = x_1 \exp(-bx_1^{-(1-m/\ell)})$, $x_1 \in (0, \varepsilon)$ is smooth and strictly monotone increasing and, hence, a homeomorphism onto some interval (p, q). Moreover, $\beta(x_1) \to 0$ as $x_1 \to 0$ and hence $p = 0$. Now $\psi : (0, q) \to (0, \varepsilon)$, the inverse of β, is well defined and $\psi(\beta) \to 0$. But by definition of $\beta(x_1)$ it follows that $x_1 = \psi(\beta)$ is the unique solution of $x_1 \ln(x_1/\beta) = bx_1^{m/\ell}$.

Replacing b by $(1/2a_\ell)^{m/\ell}$ and by $(2/a_\ell)^{m/\ell}$ we obtain φ_1 and φ_2 and hence the existence of φ follows. □

Remark 3.1. Lemma 3.1 can be modified in obvious ways to obtain a function φ for the cases where the domain of λ is $(-\varepsilon, 0]$ and/or codomain of λ is $(-\infty, 0]$. For example, if the domain of λ is $(-\varepsilon, 0]$ and the codomain is $(-\infty, 0]$, then we replace $x_1 \ln(x_1/\beta)$ by $x_1 \ln(-x_1/\beta)$ and obtain a function φ satisfying (i) and (ii) in Lemma 3.1 such that the graphs of $x_1 = \lambda(x_2)$ and $x_2 = x_1 \ln(-x_1/\beta)$ meet at $(-\varphi(\beta), -\varphi(\beta) \ln(\varphi(\beta)/\beta))$.

A second technical lemma is now needed which will be used to prove the nonexistence of nontrivial periodic orbits inside W^β.

LEMMA 3.2. *Let $\sigma \in \mathscr{C}_{r,u}$ be the graph of $\lambda : [0, \varepsilon) \to [0, \infty)$. Suppose that the region above σ does not contain critical or noncritical curves. Let $\tilde{\lambda} : [0, \varepsilon) \to [0, \infty)$ be defined by $\tilde{\lambda}(\theta) = \frac{1}{2}\lambda(\theta)$ and let $\tilde{\sigma}$ be its graph $\{(\tilde{\lambda}(\theta), \theta) \mid \theta \in [0, \varepsilon)\}$. Then for sufficiently large k the tangent vector $[f(x), -kx_2]^T$ points into the region between σ and $\tilde{\sigma}$ for all $x = (x_1, x_2)$ on $\tilde{\sigma}$.*

Proof. On $A \stackrel{\text{def}}{=} \{(x_1, x_2) \in \mathbb{R}^2_+ \mid x_2 \geqq 0\} \cap B_\varepsilon(0)$ write $f(x) = (x_1 - \lambda(x_2))\psi(x_1, x_2)$ where $\psi(x)$ is bounded. Existence of such ψ follows since $\sigma \in \mathscr{C}$. Since $\sigma \in \mathscr{C}_{r,u}$, there exists a convergent rational power series

$$\lambda(x_2) = \sum_{n \geqq \ell} a_n x_2^{n/m} \qquad (l \geqq m,\ a_\ell \neq 0).$$

If $f(x) > 0$ on $\tilde{\sigma}$, then the result is obvious. So assume that $f(x) < 0$ on $\tilde{\sigma}$. Then for x on $\tilde{\sigma}$ close enough to zero,

$$\left|\frac{f(x)}{-kx_2}\right| = \left|\frac{\psi(x_1, x_2)\lambda(x_2)}{2kx_2}\right| = \left|\frac{\psi(x)}{2k} \sum_{n \geqq \ell} a_n x_2^{(n-m)/m}\right|$$

$$\leqq \frac{L}{2k} x_2^{(l-m)/\ell} \quad \text{for some } L > 0.$$

However,

$$\frac{d}{dx_2}\tilde{\lambda}(x_2) = \sum_{n \geqq \ell} \frac{n}{m} a_n x_2^{(n-m)/m} \geqq \frac{\ell}{2m} a_\ell x_2^{(\ell-m)/\ell}.$$

Therefore, for large enough k, $f(x)/-kx_2 < (d/dx_2)\tilde{\lambda}(x_2)$ for $x \in \tilde{\sigma}$ and hence the result follows. □

The lemma below is used to prove the positive invariance of W^β.

LEMMA 3.3. *Let σ = graph of $\{\lambda : [0, \varepsilon) \to [0, \infty)\} \in \mathscr{C}_{r,u}$. Let $\beta_0 > 0$ be small enough. Let $\varphi : (0, \beta_0) \to (0, \infty)$ be as given in the conclusion of Lemma 3.1. For $\beta \in (0, \beta_0)$ let $\mu^\beta : [0, \varphi(\beta)] \to \mathbb{R}^2$ be $\mu^\beta(\theta) = (\theta, \theta \ln(\theta/\beta))$. Let A^β be the region bounded by μ^β, σ and the positive x_1-axis. Then for large enough k and for all $\beta \in (0, \beta_0)$, the tangent vector $[f(x), k(x_1 + x_2)]^T$ points into A^β for all $x = (x_1, x_2) \in \mu^\beta$.*

Proof. For each $x \in \mu^\beta$, $[x_1, x_1+x_2]^T$ is a tangent vector to μ^β. Since the graph of λ is in $\mathscr{C}$ it follows that $f(x) = (x_1 - \lambda(x_2))\psi(x)$ on A^β where $\psi(x)$ is bounded on a neighborhood of zero. Since $x_1 > \lambda(x_2)$ on A^β it follows that $f(x) < Lx_1$ for all $x \in A^\beta$ and for small enough β and for some $L > 0$. Therefore the conclusion follows at once. □

Remark 3.2. Lemmas 3.2 and 3.3 can be modified in obvious ways to incorporate the cases when $\sigma \in \mathscr{C}_{a,b}$ where $(a, b) \in \{r, \ell\} \times \{u, d\}$ to obtain conclusions similar to those already obtained. The following lemma should also be interpreted in this generalized sense.

LEMMA 3.4. *Let $\beta_0 > 0$ be small and for $\beta \in (0, \beta_0)$ let $\mu^\beta : [0, \beta] \to \mathbb{R}^2$ be the straight line joining $(0, -\beta)$ to $(\beta, 0)$ parameterized by length. Then, for large enough k, the tangent vector $[f(x), -k(x_1+x_2)]^T$ points into the region bounded by the negative x_2 axis, μ^β and the positive x_1 axis for all $x = (x_1, x_2) \in \mu^\beta$ and for all $\beta \in (0, \beta_0)$.*

Proof. Since $f(x_1, x_2)$ is C^1 (in fact C^ω), it follows that there exist $\varepsilon > 0$ and $L > 0$ such that $|f(x_1, x_2)| < L(|x_1| + |x_2|)$ for all $x \in B_\varepsilon(0)$. The result follows at once. □

Proof of Theorem 3.1. It is only necessary to prove the "if" part since the "only if" part in the theorem is obvious.

The proof is broken into several cases. The x_1 axis is already redefined, if necessary, such that none of the curves $\sigma \in \mathscr{C}$ are tangential to the x_1 axis at the origin. In particular, $f(x) \neq 0$ on the x_1 axis. Throughout the proof ε, δ, β_0, etc., are used to denote positive real numbers which are arbitrarily small without further reference to them.

Case 1. $\mathscr{C}^c = \varnothing$. Now by the hypothesis of the theorem, $x_1 f(x_1, x_2) \leqq 0$ for all $x \in B_\varepsilon(0)$. Define feedback $\alpha(x) = -x_2$, and now $\frac{1}{2}(x_1^2 + x_2^2)$ is a Lyapunov function for the closed-loop system proving the stabilizability.

Case 2. There exists a C^2 curve $\mu = (\mu_1, \mu_2) : (-\varepsilon, \varepsilon) \to \mathbb{R}^2$ such that $\mu(0) = 0$, $\mu_1(0) > 0$, and $\mu_1(0) f(\mu(0)) < 0$ for all $\theta \neq 0$.

This case can be handled trivially using center manifold theory by using the techniques outlined in Carr [10] and we will not discuss this any further.

Case 3. $\mathscr{C}^c$ contains only one element and it is in $\mathscr{C}^c_{r,u}$. Moreover, $f(x) < 0$ for x immediately above it.

Denote the element in $\mathscr{C}^c$ by σ = graph of $(\lambda : [0, \varepsilon) \to [0, \infty))$. For small β consider the open neighborhood W^β of zero shown in Fig. 3.1.

In Fig. 3.1, the curve $\tilde{\sigma}$ is defined as follows. If there exists a $\bar{\sigma} \in \mathscr{C} - \mathscr{C}^c$, which is above σ, then choose it such that there are no elements of $\mathscr{C}$ between σ and $\bar{\sigma}$ and let $\tilde{\sigma} = \bar{\sigma}$. Otherwise, let $\tilde{\lambda} : [0, \varepsilon) \to [0, \infty)$ be $\tilde{\lambda}(x_2) = \frac{1}{2}\lambda(x_2)$ and let $\tilde{\sigma}$ be the graph of $\tilde{\lambda}$. Once $\tilde{\sigma}$ is defined, the points γ_1, γ_2, and γ_3 are defined as functions of β. It follows from Lemma 3.1 that $\{W^\beta\}_{0<\beta<\beta_0}$ is a neighborhood base of the origin. S^β is defined to be the region bounded by σ, $\tilde{\sigma}$ and the curve $x_1 = \gamma_1$.

In order to define the feedback function α on $B_\varepsilon(0)$, the following open subsets of $B_\varepsilon(0) - \{0\}$ are defined. Let $0 < \omega$ be such that $(d/dx_2)\lambda(0) < \omega$.

$\mathscr{R}_1$ = Region between $\tilde{\sigma}$ and the line $x_1 = 2\omega x_2$ in $\mathbb{R}^2_+$,

$\mathscr{R}_2$ = Region between $x_1 = -\omega x_2$ and $x_1 = \omega x_2$ in $\mathbb{R}^2_+$,

$\mathscr{R}_3$ = Region between the negative x_2 axis and the line $x_1 = -2\omega x_2$ in $\mathbb{R}^2_+$,

$\mathscr{R}_4$ = Region between σ and the positive x_2 axis in $\mathbb{R}^2_+$,

$\mathscr{R}_5 = \{(x_1, x_2) \mid x_1 \leqq 0\} - \{0\}$,

$\mathscr{R}_6 = \mathscr{R}_3 \cup \mathscr{R}_4 \cup \mathscr{R}_5$.

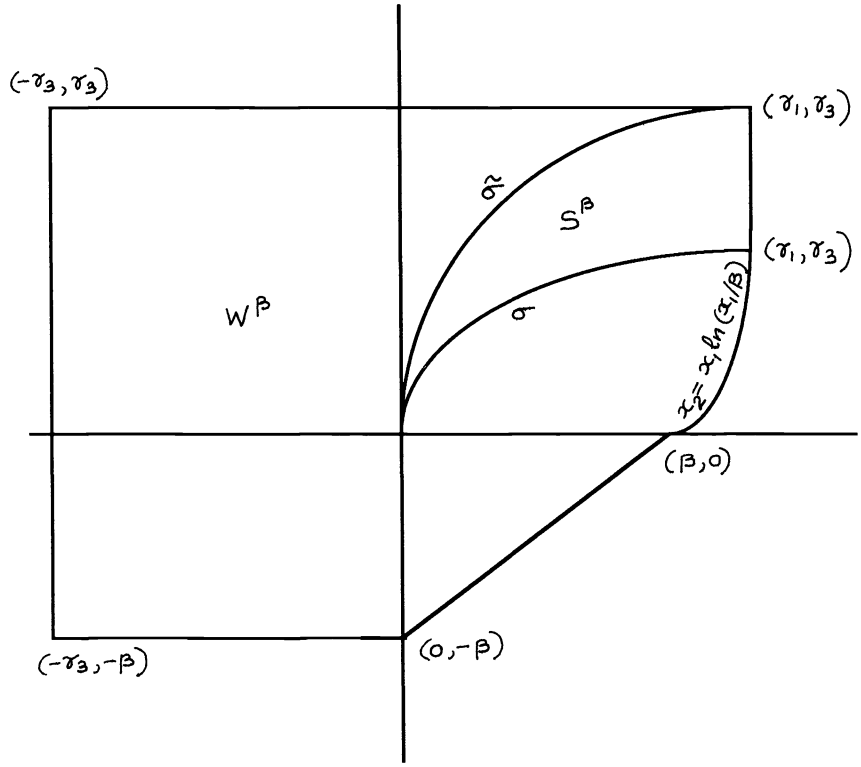


FIG. 3.1

Now $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_6$ is an open cover of $B_\varepsilon(0) = \{0\}$. Let μ_1, μ_2, μ_6 be a subordinate C^∞-partition of unity to this cover. Now define α on $B_\varepsilon(0)$ by

$$\alpha(x) = \begin{cases} 0 & \text{if } x = 0, \\ k((x_1 + x_2)\mu_1(x) + x_1\mu_2(x) - x_2\mu_6(x)) & \text{if } x \neq 0, \end{cases}$$

where k is a positive constant to be determined.

Now, since α is C^∞ on $B_\varepsilon(0) - \{0\}$ and since $|\alpha(x)| < 2k\|x\|$ on $B_\varepsilon(0)$, it follows that the closed-loop system has local unique solutions on $B_\varepsilon(0)$. We take β_0 such that $W^{\beta_0} \subset B_\varepsilon(0)$. Now it is obvious that the vector field $[f(x), \alpha(x)]^T$ points into W^β or tangential to ∂W^β at all x on the horizontal and the vertical parts of the boundary. By Lemmas 3.3 and 3.4 the same conclusion holds on $x_2 = x_1 \ln(x_1/\beta)$ and on $x_1 + x_2 = \beta$ for all $\beta < \beta_0$ whenever k is large enough. So for such k, W^β is positively invariant for all $\beta < \beta_0$. Since zero is the only equilibrium point of the closed-loop system in W^β $(\beta \in (0, \beta_0))$, it follows from the Poincaré–Bendixon theorem [17] that nontrivial periodic orbits of the closed-loop system in W^β (if any) should encircle the origin. However, S^β is obviously positively invariant; therefore, we conclude that there are no nontrivial periodic orbits in W^β at all. Hence, by the Poincaré–Bendixon theorem for continuous vector fields [17], it follows that for each $x^0 \in W^\beta$ the positive limit set $\omega(x^0)$ is the origin. Since $\{W^\beta\}_{\beta < \beta_0}$ is a neighborhood base, the asymptotic stability of the closed-loop system follows.

The argument for proving the asymptotic stability of the closed-loop system is essentially the same for all of the remaining cases. Therefore, we will only describe the construction of W^β and α for those cases and omit writing the argument.

Case 4. $\mathscr{C}^c$ contains only one element and it is in $\mathscr{C}^c_{r,u}$. Moreover, $f(x)<0$ is immediately below it.

The element in $\mathscr{C}^c$ is denoted by $\sigma=$ graph of $\lambda:[0,\varepsilon)\to[0,\infty)$. Before describing W^β, a C^1 function $\tilde{\lambda}:[0,\varepsilon)\to[0,\infty)$ will be defined in the following way. If there exist some $\hat{\sigma}=$ graph of $\hat{\lambda}:[0,\varepsilon)\to[0,\infty)$ in $\mathscr{C}$ below σ and above the line $x_2=0$, then the uppermost such $\hat{\sigma}$ is taken and we define $\tilde{\lambda}$ to be $\hat{\lambda}$. Otherwise we define $\tilde{\lambda}=2\lambda$. Now $\tilde{\sigma}$ will denote the graph of $\tilde{\lambda}$. For small enough β we will now define W^β to be the region shown in Fig. 3.2. The point (γ_1,γ_2) is the point of intersection of $x_2=x_1\ln(x_1/\beta)$ and $x_1=\tilde{\lambda}(x_2)$. (This point exists by Lemma 3.1 and Remark 3.1 and $(\gamma_1,\gamma_2)\to 0$ as $\beta\to 0$.) The point (γ_1,γ_3) is the point of intersection of $x_1=\gamma_1$ and $x_1=\lambda(x_2)$. The region S^β is the region enclosed by $\tilde{\sigma}$, σ and $x_1=\gamma_1$.

In order to define the feedback function α the following regions are defined. Let $\omega>0$ be such that $(d/dx_2)\tilde{\lambda}(0)<\omega$.

$$\mathscr{R}_1=\{(x_1,x_2)\in B_\varepsilon(0)-\{0\}\mid x_1\leqq 0\}$$
$$\cup\{(x_1,x_2)\in B_\varepsilon(0)\cap\mathbb{R}^2_+\mid x_2>0,\ 0<x_1<\tilde{\lambda}(x)\}.$$
$$\cup\{(x_1,x_2)\in B_\varepsilon(0)\cap\mathbb{R}^2_+\mid x_2<0,\ 0<x_1<2x_2\},$$
$$\mathscr{R}_2=\left\{(x_1,x_2)\in B_\varepsilon(0)\cap\mathbb{R}^2_+\;\middle|\;-x_1<x_2<\frac{1}{\omega}x_1\right\},$$
$$\mathscr{R}_3=\{(x_1,x_2)\in B_\varepsilon(0)\cap\mathbb{R}^2_+\mid x_2>0,\ \lambda(x_2)<x_1<2\omega x_2\}.$$

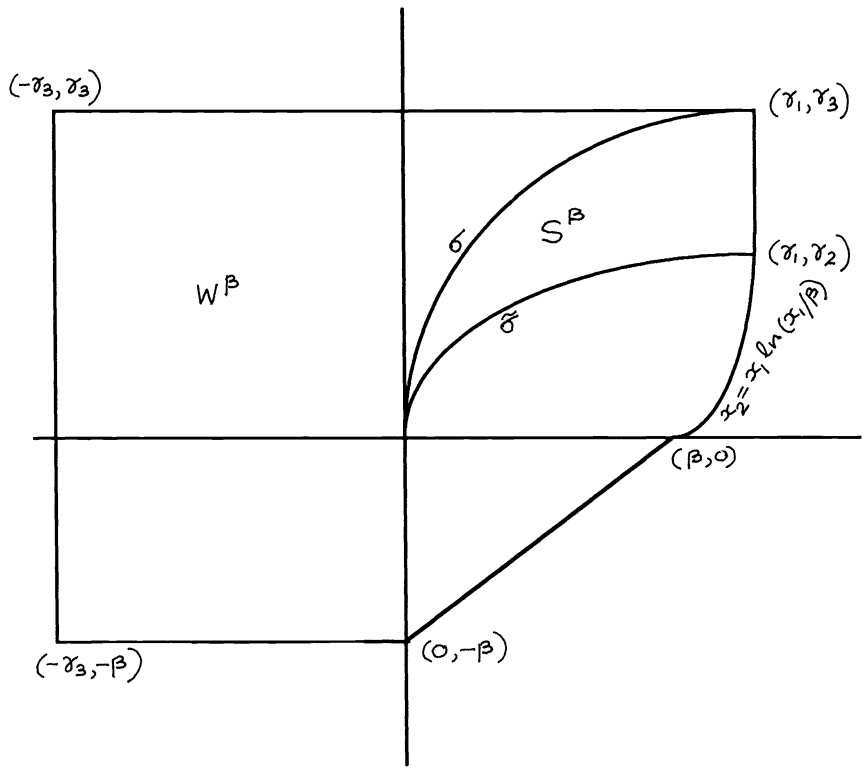


FIG. 3.2

Let μ_1, μ_2, μ_3 be a C^∞-partition of unity of $B_\varepsilon(0)-\{0\}$ subordinate to the open cover $\bigcup_{i=1}^{3}\mathcal{R}_i$. Now define the feedback function as

$$\alpha(x)=\begin{cases}0 & \text{if } x=0,\\ k(-x_2\mu_1(x)+x_1\mu_2(x)+x_2\mu_3(x)), & \end{cases}$$

where k is a large positive constant.

The following facts are now easily verified.

(1) Let β_0 be small enough and let k be large enough. Then $\{W_\beta\}_{\beta<\beta_0}$ is a neighborhood base of the origin and each W_β and S_β is positively invariant for the closed-loop system.

(2) The only equilibrium point of the closed-loop system in W^β is the origin.

Therefore, by arguing as in the previous case, the asymptotic stability of the closed-loop system is proven.

Case 5. There exist $\sigma\in\mathscr{C}^c_{r,u}$ and $\nu\in\mathscr{C}^c_{\ell,u}$. Moreover, $f(x)<0$ for all x immediately above ν and $f(x)>0$ for all x immediately above ν.

Define a C^1 curve $\tilde{\sigma}$ below σ and a C^1 curve $\tilde{\nu}$ below ν as was done in Case 3. Now for small enough β, W^β is defined to be the neighborhood of the origin shown in Fig. 3.3, and S^β will denote the region enclosed by σ, $\tilde{\sigma}$ and the curve $x_1=\gamma_1$ in Fig. 3.3.

Let $\omega>0$ be such that the line $x_1=\omega x_2$ is below $\tilde{\sigma}$ and $x_1=-\omega x_2$ is below $\tilde{\nu}$.

Let

$$\mathcal{R}_1=\text{Region above } \tilde{\nu}\cup\tilde{\sigma} \text{ in } B_\varepsilon(0)-\{0\},$$

$$\mathcal{R}_2=\text{Region between } \sigma \text{ and } x_1=2\omega x_2 \text{ in } \mathbb{R}^2_+\cap(B_\varepsilon(0)-\{0\}),$$

$$\mathcal{R}_3=\text{Region between } \nu \text{ and } x_1=-2\omega x_2 \text{ in } \mathbb{R}^2_-\cap(B_\varepsilon(0)-\{0\}),$$

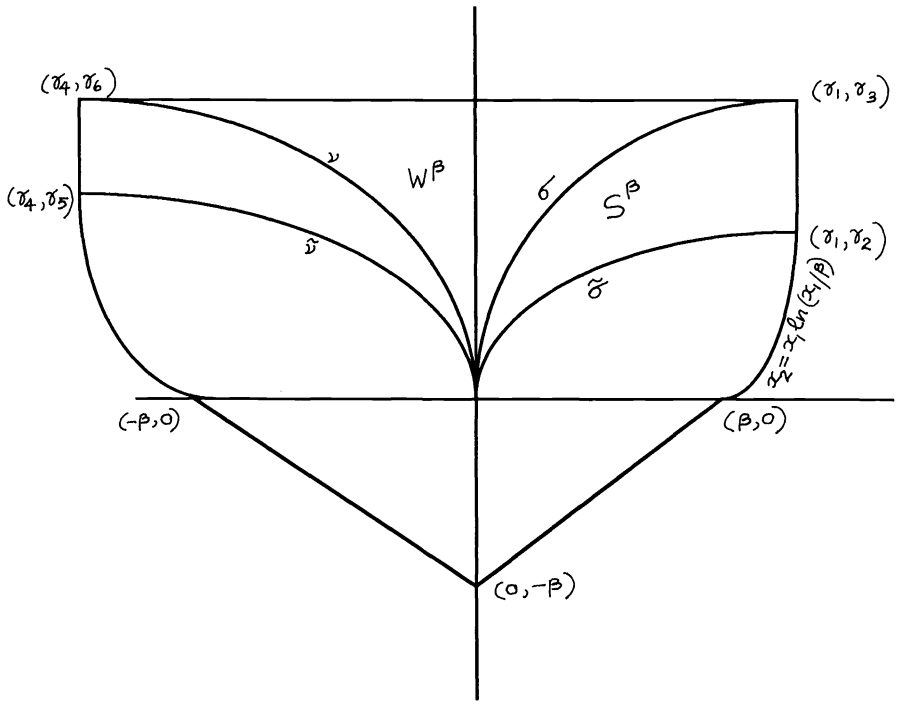


FIG. 3.3

$$\mathcal{R}_4 = \text{Region between } x_1 = \omega x_2 \text{ and } x_1 = \omega x_2 \text{ in } B_\varepsilon(0) - \{0\},$$

$$\mathcal{R}_5 = \text{Region below } x_1 = -2\omega x_2 \text{ in } \mathbb{R}^2_+ \cap (B_\varepsilon(0) - \{0\})$$

$$\cap \text{ Region below } x_1 = 2\omega x_2 \text{ in } \mathbb{R}^2_- \cap (B_\varepsilon(0) - \{0\}).$$

Let $\{\mu_i\}_{i=1}^5$ be a C^∞-partition of unity on $B_\varepsilon(0) - \{0\}$ subordinate to the open cover $\cup_{i=1}^5 \mathcal{R}_2$. Now define $\alpha : B_\varepsilon(0) \to \mathbb{R}$ by

$$\alpha(x) = \begin{cases} 0, & x = 0, \\ k(-x_2\mu_1(x) + x_2\mu_2(x) + x_2\mu_3(x) + x_1\mu_4(x) - x_1\mu_5(x)), \end{cases}$$

where k is a large positive constant.

It is easily proven that there exist large enough k and small enough $\beta_0 > 0$ such that $\{W^\beta\}_{\beta<\beta_0}$ and $\{S^\beta\}_{\beta<\beta_0}$ are positively invariant for the closed-loop system, they contain no nontrivial periodic orbits, and the origin is the only equilibrium point in W^β for all $\beta < \beta_0$. Therefore, the asymptotic stability follows.

Case 6. There exist $\sigma \in \mathcal{C}^c_{r,u}$ and $\nu \in \mathcal{C}^c_{\ell,u}$ and $f(x) > 0$ for all x immediately above σ and $f(x) < 0$ for all x immediately above ν.

Now consider $\tilde{\sigma}$ above σ and $\tilde{\nu}$ above ν as was done in Case 4. In this case the regions $\{\mathcal{R}_i\}_{i=1}^5$ are constructed as in Case 5 but interchange σ and $\tilde{\sigma}$ and ν and $\tilde{\nu}$. Now construct α using the formula in Case 5 and argue as in Case 3 to prove asymptotic stability.

Case 7. There exist $\sigma \in \mathcal{C}^c_{r,u}$ and $\nu \in \mathcal{C}^c_{\ell,u}$ and $f(x) > 0$ for all x immediately above $\sigma \cup \nu$.

Now construct $\tilde{\sigma}$ above σ as in Case 4 and $\tilde{\nu}$ below ν as in Case 3. Define the regions $\{\mathcal{R}_i\}_{i=1}^5$ as in Case 5 but interchange σ and $\tilde{\sigma}$.

Case 8. There exist $\sigma \in \mathcal{C}^c_{r,u}$ and $\nu \in \mathcal{C}^c_{\ell,d}$. Moreover, $f(x) > 0$ for all x immediately above $\sigma \cup \nu$.

Let $\sigma = \text{graph of } \lambda : [0, \varepsilon) \to [0, \infty)$ and $\nu \neq \text{graph of } \nu : (-\varepsilon, 0] \to (-\infty, 0]$. Define $\tilde{\lambda} : [0, \varepsilon) \to (0, \infty)$ as in Case 4. If there are no elements of $\mathcal{C}_{\ell,d}$ above ν, then define $\tilde{\nu} : (-\varepsilon, 0] \to (-\infty, 0]$ by $\tilde{\nu}(x_2) = 2\nu(x_2)$ and let $\tilde{\nu}$ be the graph of $\tilde{\nu}$. Otherwise let $\tilde{\nu}$ be the element of $\mathcal{C}_{\ell,d}$, which is immediately above ν. Now for small enough $\beta > 0$ define the region W^β as shown in Fig. 3.4.

Let S^β be the region bounded by σ, $\tilde{\sigma}$ and the line $x_1 = \gamma_1$. Let $\omega > 0$ be such that the line $x_1 = \omega x_2$ is below $\tilde{\sigma}$ and above $\tilde{\nu}$.

Let $\mathcal{R}_1$ be the region above $\tilde{\sigma} \cup \{(x_1, x_2) \mid x < 0, x_2 + (1/2\omega)x_1\}$. Let $\mathcal{R}_2$ be the region below $x_1 = -\omega x_2$ $(x_1 < 0)$ and above $x_1 = \omega x_2$ $(x_1 < 0)$. Let $\mathcal{R}_3$ be the region below $x_1 = 2\omega_2$ $(x_1 < 0)$ and above ν.

Let $\mathcal{R}_4$ be the region between $\tilde{\nu}$ and $x_1 = 2\omega x_2$ $(x_1 > 0)$. Let $\mathcal{R}_5$ be the region above $x_1 = -\omega x_2$ $(x_1 > 0)$ and below $x_1 = \omega x_2$ $(x_1 > 0)$. Let $\mathcal{R}_6$ be the region above $x_1 = 2\omega x_2$ $(x_1 > 0)$ and below σ.

Let $\{\mu_i\}_{i=1}^6$ be a C^∞-partition of unity subordinate to the cover $\cup_{i=1}^6 \mathcal{R}_i$ of $B_\varepsilon(0) - \{0\}$. Now define the feedback function $\alpha : B_\varepsilon(0) \to \mathbb{R}$ by

$$\alpha(x) = \begin{cases} 0, & x = 0, \\ k(-x_2\mu_1(x) + x_1\mu_2(x) + x_2\mu_3(x) - x_2\mu_4(x) + x_1\mu_5(x) + x_2\mu_6(x)), \end{cases}$$

where k is a large positive constant.

Now, as in Case 3, it is argued that when k is large enough, and β_0 is small enough, $\{W^\beta\}_{\beta<\beta_0}$ and $\{S^\beta\}_{\beta<\beta_0}$ are all positively invariant, $\{W^\beta\}_{\beta<\beta_0}$ does not contain nontrivial periodic orbits, and the origin is the only equilibrium point of the closed-loop system in W^β. Since $\{W^\beta\}_{\beta<\beta_0}$ is a neighborhood base of the origin, the asymptotic stability follows at once.

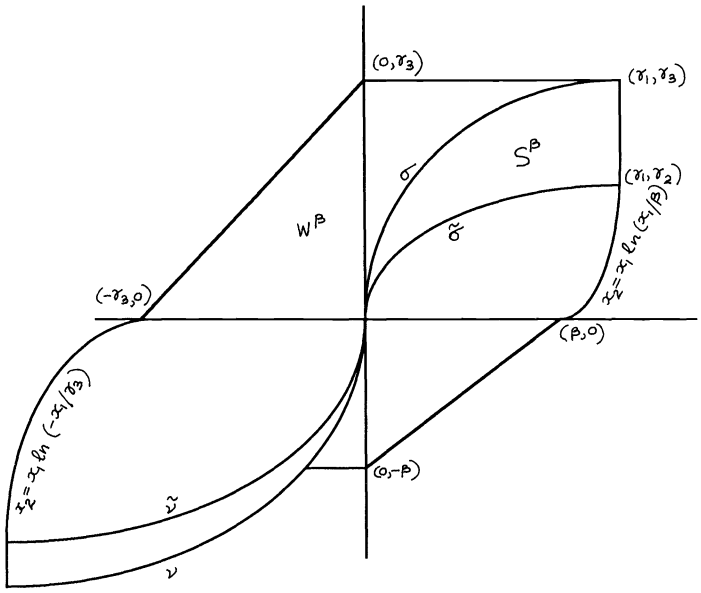


FIG. 3.4

Case 9. There is $\sigma \in \mathscr{C}^c_{r,u}$, $\nu \in \mathscr{C}^c_{\ell,d}$, and $f(x) > 0$ for x immediately above σ or immediately below ν. Now define $\tilde{\sigma}$ as in Case 8 and $\tilde{\nu}$ above ν to be the element of $\mathscr{C}_{\ell,d}$ immediately below ν if such an element exists or let $\tilde{\nu} = \{(x_1, x_2) \mid x_2 \in (-\varepsilon, 0], (2x_1, x_2) \in \nu\}$.

Now define $\{\mathscr{R}_i\}_{i=1}^6$ as in Case 8, but interchange ν and $\tilde{\nu}$ and define α accordingly and complete the argument as in Case 8.

Each of the remaining cases is essentially similar to a case considered already in the sense that the modifications needed are obvious.

4. Stabilization of real analytic systems using smooth feedback. Consider a two-dimensional real analytic control system,

$$\dot{x}_1 = \tilde{f}(x_1, x_2), \qquad \dot{x}_2 = u, \tag{4.1}$$

where u is the control and $\tilde{f}$ is a real analytic function defined near the origin of $\mathbb{R}^2$ and $f(0) = 0$. As seen before, there is no loss in generality in considering the special case

$$\dot{x}_1 = f(x_1, x_2), \qquad \dot{x}_2 = u, \tag{4.2}$$

where $f(x)$ is a Weierstrass polynomial. Let

$$f(x) = x_1^m + a_1(x_2)x_1^{m-1} + \cdots + a_m(x_2).$$

The objective in this section is to give a rather sharp sufficient condition for feedback stabilization of (4.1) using C^1-feedback and to derive some obstructions for the existence of C^∞-stabilizing feedback. It is said that (4.1) *is* C^r *stabilizable if there exist* C^r *feedback which asymptotically stabilizes the system.*

Throughout this section the notation introduced in the previous section regarding the variety $f^{-1}(0)$ will be used.

In what follows, we will show that there are three indices which seem to dictate the smoothness of stabilizing feedback. The first of them quantifies the vertical distance that a trajectory should travel before its horizontal component begins to move towards

the origin. The remaining quantifiers dictate the smallest value of the feedback function needed in order to cross certain critical boundary curves. It turns out that in certain cases even the smallest value is too large in the sense that to achieve it the feedback function must vary too rapidly. Therefore, such cases point out obstructions for smooth feedback stabilizability.

Throughout this section ε will denote a positive real number which is as small as desired.

Denote the positive rationals by Q_+ and define:

$$A^+ = \{\gamma \in Q_+ \mid f(x_1, \phi(x_1)) < 0 \text{ for all } x_1 \in (0, \varepsilon), \text{ for some } \varepsilon > 0, \text{ and for some convergent rational power series } \phi(x_1) \text{ with leading exponent equal to } 1/\gamma\},$$

$$A^- = \{\gamma \in Q_+ \mid f(-x_1, \phi(x_1)) > 0 \text{ for all } x_1 \in (0, \varepsilon), \text{ for some } \varepsilon > 0 \text{ and for some convergent rational power series } \phi(x_1) \text{ with leading exponent equal to } 1/\gamma\}.$$

DEFINITION 4.1. The index of stabilizability of f is $\max\{\inf_{\gamma \in A^+}\{\gamma\}, \inf_{\gamma \in A^-}\{\gamma\}\}$.

DEFINITION 4.2. The fundamental stabilizability degree of f is the order of the zero of $a_m(x_2)$ at $x_2 = 0$. The secondary stabilizability degree of f is the order of the zero of $a_{m-1}(x_2)$ at $x_2 = 0$.

Notation.

$I :=$ Index of stabilizability of f,

$s_1 :=$ Fundamental stabilizability degree of f,

$s_2 :=$ Secondary stabilizability degree of f.

Note that s_1 is invariant under real analytic weak feedback equivalence. Indeed we can define s_1 directly from $\tilde{f}$ as the smallest integer ℓ such that

$$\frac{\partial^\ell \tilde{f}}{\partial x_2^\ell}(0) \neq 0.$$

The remaining indices are not weak feedback invariants in general.

Note that according to Theorem 3.1, I is defined only when (4.2) is C^0-stabilizable.

In what follows, we need to compare s_1 with $2I-1$ and s_2 with $I-1$ when $I > 0$. By considering the factorization,

$$f(x) = (x_1 - \gamma_1(x_2))(x_1^{m-1} + c_1(x_2)x_1^{m-2} + \cdots + c_{m-1}(x_2))$$

where $\gamma_1(x_2)$ is a rational power series with leading exponent equal to I and $c_i(x_2)$, $i = 1, \cdots, m-1$ are rational power series in x_2, it is easily concluded that $s_2 > I-1$ (respectively, $s_2 \geqq I-1$) if $s_1 > 2I-1$ (respectively, $s_1 \geqq 2I-1$).

THEOREM 4.1. *The system* (4.2) *and hence* (4.1) *is* C^1*-stabilizable if*

$$s_1 > 2I - 1.$$

Proof. If $I = 0$, then obviously $\alpha(x) = -kx_2$ stabilizes the system for large positive k. When $I \neq 0$, the proof of this theorem is very much like the proof of Theorem 3.1 on C^0-stabilizability except for the construction of the feedback function. The invariant sets $\{W^\beta\}_{\beta < \beta_0}$ will be exactly the same as before and the proof of the invariance follows along the same lines. So here we will only construct $\alpha(x)$ in a representative case and leave the details to the reader.

For convenience we will assume that there exist $\sigma_r \in \mathscr{C}^c_{r,u}$, which is the graph of $x_1 = \lambda_r(x_2)$; $\sigma_\ell \in \mathscr{C}^c_{\ell,u}$, which is the graph of $x_1 = \lambda_\ell(x_2)$; and that the leading exponents γ_r and γ_ℓ in the rational power series of λ_r and λ_l are less than or equal to I. Moreover, it is assumed that $f(x) < 0$ above σ_r, $f(x) > 0$ above σ_ℓ, that there are elements of $\mathscr{C}_r$ above σ_r with the leading exponent γ_r, and no elements of $\mathscr{C}_\ell$ are above σ_l with the leading exponent γ_l. Define $\nu_r : [0, \varepsilon) \to [0, \infty)$ and $\nu_\ell : [0, \varepsilon) \to [0, \infty)$ by $\nu_r(x_2) = \frac{1}{2}\lambda_r(x_2)$ and $\nu_l(x_2) = \frac{1}{2}\lambda_l(x_2)$. (If there are elements of $\mathscr{C}_r$ or $\mathscr{C}_\ell$ above σ_r or σ_l with leading exponents γ_r and γ_ℓ, respectively, then we take ν_r and/or ν_ℓ to be some rational power series with leading exponents γ_r and/or γ_ℓ and such that their graphs lie between σ_r and/or σ_ℓ and the corresponding element of C_r of C_{ℓ^*}) Now ν_v and ν_ℓ are strictly monotone rational power series and, therefore, we can invert $x_1 = \nu_a(x_2)$ $(a \in \{r, \ell\})$ to obtain

$$\eta_r : [0, \varepsilon) \to [0, \infty) \quad \text{and} \quad \eta_\ell : (-\varepsilon, 0] \to [0, \infty),$$

which are both convergent rational power series and σ_r and σ_ℓ are the groups of $x_2 = \eta_r(x_1)$ and $x_2 = \eta_\ell(x_1)$, respectively. Now the leading coefficients of η_r and η_ℓ are $1/\gamma_r$ and $1/\gamma_\ell$, respectively, and each is not less than l/I. Now let k be a large positive constant and γ is a rational of the form $(2p+1)/(2q+1)$ for $p, q \in \mathbb{N}$. If $I \leqq 1$, γ is taken to be greater than but very close to $1/I$ and if $I > 1$, we take $\gamma = 1$. Now define the feedback function $\alpha(x)$ as

$$\alpha(x) = \begin{cases} 0 & \text{if } x = 0, \\ -kx_2^\gamma + k(\eta_r(x_1))^\gamma & \text{for } x_1 > 0, \\ -kx_2^\gamma + k(\eta_\ell(x_1))^\gamma & \text{for } x_1 > 0. \end{cases}$$

It is obvious that α is C^1.

The sets $\{W^\beta\}_{\beta < \beta_0}$ are constructed now as in the proof of Theorem 3.1. The positive invariance of these sets and the nonexistence of any equilibrium points other than the one at the origin is verified exactly along the same lines as before using the inequalities $s_1 > 2I - 1$ and $s_2 > I - 1$. These inequalities permit $\alpha(x)$ to be made much smaller than those we had constructed in Theorem 3.1 and, thereby, produce C^1 feedback functions. The rest of the details are left to the interested reader. □

COROLLARY 4.1. *Suppose that f is symmetric with respect to x_1 (i.e., $f(x_1, x_2) = f(-x_1, x_2)$). Then f is C^1 stabilizable if and only if f is C^0 stabilizable.*

Proof. It is only needed to show that C^0-stabilizability implies C^1-stabilizability. C^0-stabilizability implies that I is defined. Since the case $m = 0$ is trivial it will be assumed that $m \geqq 2$.

The following cases are possible.

Case 1. $I = 0$. Now the symmetry of f implies that $f = 0$ on the x_1-axis. Therefore, $s_1 = \infty$ and the desired conclusion follows from Theorem 4.1.

Case 2. $I \neq 0$. Now we have a convergent rational power series $\lambda : \mathscr{U}_\varepsilon \to \mathbb{R}$ with leading exponent I such that

$$f(\lambda(x_2), x_2) = 0 \quad \text{for all } x_2 \in \mathscr{U}_\varepsilon,$$

where $\mathscr{U}_\varepsilon$ is either $[0, \varepsilon)$ or $(-\varepsilon, 0]$.

But, by symmetry of f with respect to x_1, it follows that

$$f(-\lambda(x_2), x_2) = 0 \quad \text{for } x \in \mathscr{U}_\varepsilon \text{ also.}$$

Therefore, $x_1^2 - (\lambda(x_2))^2$ is a factor of $f(x_1, x_2)$ for $x_2 \in \mathscr{U}_\varepsilon$.

It now follows that the leading exponent of $a_m(x_2)$ (i.e., s_1) is not less than the leading exponent of $(\lambda(x_2))^2$.

Therefore, $s_1 \geqq 2I$. It follows from Theorem 4.1 that the system is C^1-stabilizable. □

In the proof of the following theorem the technique for the construction of the Lyapunov function is due to Kawski [18].

THEOREM 4.2. *Suppose that* $1+2s_2 \geqq s_1$ *and* s_1 *is odd. Then the system* (4.1) *is* C^ω*-stabilizable.*

Proof. Since the proof is trivial if $m=0$ we assume that $m>1$.

By replacing x_2 by $-x_2$ if necessary we may assume that the leading coefficient of $a_m(x_2)$ is positive. Let $a_i(x_2)=\sum_{j=1}^{\infty} a_{i,j}x_2^j$. Now let $k>0$ be large enough and define the function

$$V(x)=x_1^m x_2+\sum_{i=1}^{m-2} x_1^{(m-i)} \sum_{j+1}^{\infty} \frac{a_{i,j}}{j=1} x_2^{j+1}+x_1 \sum_{j=s_2}^{\infty} \frac{a_{m-1,j}}{j+1} x_2^{j+1}$$

$$+\sum_{j=s_1+1}^{\infty} \frac{a_{m,j}}{j+1} x_2^{j+1}+\frac{a_{m,s_1}}{s_1+1} x_2^{s_1+1}+kx_1^2.$$

Now define the feedback function $\alpha(x)=-\partial V/\partial x_1-\partial V/\partial x_2$. Obviously, α is real analytic.

Since s_1 is odd and $s_2+1 \geqq \frac{1}{2}(s_1+1)$ it follows that for large enough values of k, the function V is positive definite on a small neighborhood of the origin. Furthermore,

$$\dot{V}(x)=-\left(\frac{\partial V}{\partial x_2}(x)\right)^2.$$

If $\dot{V}(x(t))\equiv 0$ along a trajectory $x(t)$ of the closed-loop system, then (since $f(x)=\partial V/\partial x_2$) it follows that $x_1(t)$ is constant.

Now rewrite $V(x)=\sum_{n=1}^{\infty} b_n(x_1)x_2^{n-1}$. Then

$$\frac{\partial V}{\partial x_2}=\sum_{n=1}^{\infty} n.b._n(x_1)x_2^{n-1},$$

so if $\partial v/\partial x_2(x(t))\equiv 0$, but $x(t)$ is nonconstant, then it follows that $b_n(x_1(t))=0$ for $n=1,2,\cdots$.

Since $b_{s_1+1}(x_1)$ is nonzero and real analytic, it follows that there exists $\varepsilon>0$ such that $b_{s_1+1}(x_1)\neq 0$ for all $0<|x_1|<\varepsilon$. Therefore, for small enough $\delta>0$ if $\{x(t),\, t>0\}\subset B_\delta(0)$ and if $V(x(t))=0$ for all $t>0$, then $x(t)$ is constant. We will now show that for possibly smaller values of δ this implies that $x(t)\equiv 0$. For if $x(t)$ is constant, then $(\partial V/\partial x_1)(x(t))=0=(\partial V/\partial x_2)(x(t))$. Since $(\partial/\partial x_1)(\partial V/\partial x_1)(0)=2k\neq 0$ it follows that $(\partial V/\partial x_1)^{-1}(0)$ is the graph of a real analytic function $x_1=\varphi(x_2)$. However, $\partial V/\partial x_2$ does not depend on k and hence for almost all values of k and for small δ, $(\partial V/\partial x_1)^{-1}(0)$ and $(\partial V/\partial x_2)^{-1}(0)$ intersect in $B_\delta(0)$ only at the origin. This now shows that for large enough k and small enough δ, $\dot{V}(x(t))\equiv 0$ and $x(t)\in B_\delta(0)$ for all $t>0$ implies that $x(t)\equiv 0$, and hence the stability follows from LaSalle's theorem. □

Now our attention is focused on obtaining necessary conditions for smooth stabilization. The result shows that if we reverse at least one of the inequalities in Theorem 4.3 then we have an obstruction to C^∞-stabilizability.

THEOREM 4.3. *Suppose that* $s_1<2I-1$. *Then the system* (4.1) *is not* C^∞*-stabilizable.*

Proof. Since s_1 is a natural number it follows that $I>1$. By the definition of I one of the following should hold.

(a) There is a sector $\mathcal{R}_1$ in $\mathbb{R}^2_+$ bounded by curves $x_1=\lambda_1(x_2)$ $(x_2\in[0,\varepsilon))$ and $x_1=\lambda_2(x_2)$ $(x_2\in(\varepsilon,0])$, such that graphs of λ_1 and λ_2 are both in $\mathcal{C}_r$ and $f(x)>0$ on $\mathcal{R}_1$, and such that the leading exponents of λ_1 and λ_2 are not less than I.

(b) There is a sector $\mathcal{R}_2$ and $\mathbb{R}^2_-$ bounded by curves $x_1=\lambda_3(x_2)$ $(x_2\in[0,\varepsilon))$ and $x_1=\lambda_4(x_2)$ $(x_2\in(-\varepsilon,0])$, such that graphs of λ_3 and λ_4 are both in $\mathcal{C}_\ell$ and $f(x)<0$ on $\mathcal{R}_2$, and such that the leading exponents of λ_3 and λ_4 are not less than I.

(c) Same as in (a) except for that either $\lambda_1(x_2)=0$ or $\lambda_2(x_2)=0$, but not both.

(d) Same as in (b) except for that either $\lambda_3(x_2)=0$ or $\lambda_4(x_2)=0$, but not both.

Now, suppose that the system is C^∞-stabilizable with feedback $\alpha(x)$. Let $\tilde{x}$ be a point close to the origin on the positive x_1-axis and let $\hat{x}$ be a point close to the origin on the negative x_1-axis. Let $x(t;\tilde{x})$ and $x(t,\hat{x})$ $(t\geqq 0)$ be the positive orbits of $\tilde{x}$ and $\hat{x}$, respectively. Let $F(x)=[f(x),\alpha(x)]^T$. Since $x(t,\tilde{x})\to 0$ and $x(t,\hat{x})\to 0$ as $t\to\infty$ the following hold.

In cases (a) and (c), $x(t;\tilde{x})$ crosses one of the boundary curves of $\mathcal{R}_1$ and leaves $\mathcal{R}_1$, and, until $x(t;\tilde{x})$ reaches the boundary of $\mathcal{R}_1$, the x_1 coordinate of $x(t,\tilde{x})$ increases monotonically.

In cases (b) and (c), $x(t;\hat{x})$ crosses one of the boundary curves of $\mathcal{R}_2$ and leaves $\mathcal{R}_2$, and, until $x(t;\hat{x})$ reaches the boundary of $\mathcal{R}_2$, the x_1 coordinate of $x(t;\hat{x})$ decreases monotonically.

For convenience, it is appropriate to only consider the cases (a) and (c), to assume that $x_1=\lambda_1(x_2)$ $(x_2\in[0,\varepsilon))$ is a boundary of $\mathcal{R}_1$, and to assume that the vector field $F(x)$ points away from $\mathcal{R}_1$ on $x_1=\lambda_1(x_2)$. Moreover, we assume that there is a sequence of points $\{\tilde{x}^n\}_{n=1}^\infty$ on the positive x_1-axis such that $\tilde{x}^n\to 0$ as $n\to\infty$ and $x(t,\tilde{x}^n)$ crosses $x_1=\lambda_1(x_2)$ and leaves $\mathcal{R}_1$ in positive time for each n. The argument given for this case is representative of the argument needed to obtain the desired contradiction in each of the remaining cases.

Now let ν be the leading exponent of $\lambda_1(x_2)$. Then $\nu\geqq I$. Now for all large $a>0$, $x(t,\tilde{x}^n)$ should cross the curve $x_1=ax_2^I$ $(x_2\in[0,\varepsilon))$ for each n. Therefore, we have a sequence $\{x^n\}_{n=1}^\infty$ on $x_1=ax_2^I$ $(x_2\in[0,\varepsilon))$ such that $x^n\to 0$ as $n\to\infty$ and

$$\frac{f(x^n)}{\alpha(x^n)}<aI(x_2^n)^{I-1} \quad \text{for all } n.$$

Since $s_1<2I-1$, it follows that for suitably large values of a,

$$f(x^n)=f(a(x_2^n)^I,x_2)\geqq c(x_2^n)^\beta \quad \text{for some } c>0 \quad \text{where } \beta=\min\{s_1,s_2+I\}.$$

Therefore,

$$\alpha(x^n)>\tilde{b}(x_2^n)^n \quad \text{for some constant } \tilde{b}>0 \quad \text{where } \eta=\beta+1-I.$$

Note that $\eta<I$. Therefore, it follows that

$$\frac{\partial^l}{\partial x_2^\ell}\alpha(0)\neq 0 \quad \text{for some } \ell\leqq\eta.$$

Therefore, for large enough a and small enough $\varepsilon>0$, there exists a positive constant b such that

$$\alpha(ax_2^I,x_2)>b(x_2)^n \quad \text{for all } x_2\in[0,\varepsilon).$$

Now let θ be a positive real number and consider the sector $\mathcal{R}_3$ bounded by $x_1=\lambda_1(x_2)$ and $x_1=-\theta x_2^I$, $x_2\in[0,\varepsilon)$. It is claimed that for some $\theta>0$ one of the following should hold: Either there is a sequence $\{x^n\}_{n=1}^\infty$ on $x_1=ax_2^I$ such that $x^n\to 0$ as $n\to\infty$ and on each line segment $\ell_n\equiv x_2=x_2^n$ in $\mathcal{R}_3$ there is a point at which $\alpha(x)<0$,

or there is a sequence $\{x^n\}_{n=1}^{\infty}$ on $x_1=-\theta x_2^I$ such that $x_2^n\to 0$ as $n\to\infty$ and $\alpha(x^n)<(b/2)(x_2^n)^\eta$ for all n. For suppose that for all $\theta>0$ there is some $\varepsilon>0$ and $\alpha(x)>(b/2)(x_2)^\eta$ on $x_1=-\theta x_2^I$ for all $x_2\in[0,\varepsilon)$. Then,

$$\frac{1}{I\theta x_2^{I-1}}\frac{|f(-\theta x_2^I,x_2)|}{\alpha(-\theta x_2^I,x_2)}\leqq\frac{2|f(-\theta x_2^I,x_2)|}{|I\theta x_2^{I-1}(b(x_2)^\eta)|}$$

$$\leqq\frac{2|d_1\theta x_2^{I+s_2}+d_2x_2^{s_1}|}{Ib\theta x_2^\beta}\quad\text{for } x_2\in[0,\varepsilon),$$

where d_1 and d_2 are positive constants which are independent of θ.

If $s_1\leqq s_2+I$ it is obvious that we can make

$$2\frac{|d_1\theta x_2^{I+s_2}+d_2x_2^{s_1}|}{Ib\theta x_2^\beta}$$

less than one by making θ large enough. If $s_1>s_2+I$, then by keeping track of the bounds obtained in proving that $\alpha(ax_2^I,x_2)>b(x_2)^\eta$, it is observed that

$$2\frac{|d_1ax_2^{I+s_2}+d_2x_2^{s_1}|}{Ibax_2^\beta}<1\quad\text{for } x_2\in[0,\varepsilon).$$

Hence, in either case the vector field F points into the sector $\mathcal{R}_3$ on the boundary curve $x_1=-\theta x_2^I$ for some $\theta>0$. Now, since the same holds true on $x_1=\lambda_1(x_2)$, it follows that $\mathcal{R}_3$ is positively invariant. But now asymptotic stability implies that each horizontal line segment ℓ in $\mathcal{R}_3$ with endpoints on boundary curves should have a point at which α is negative. Therefore, the claim has been proved.

Now let $\theta>0$ and $\mathcal{R}$ be as in the claim above. Then, it follows from the claim that there exists a sequence $\{x^n=(x_1^n,x_2^n)\}_{n=1}^{\infty}$ in $\mathcal{R}$ converging to the origin such that $\alpha(x^n)<(b/2)(x_2^n)^\eta$. Then

$$\frac{|\alpha(a(x_2^n)^I,x_2^n)-\alpha(x^n)|}{|(a(x_2^n)^I,x_2^n)-x^n|}\geqq\frac{b}{2}\frac{(x_2^n)^\eta}{(a+\theta)(x_2^n)^I}=\frac{b}{2(a+\theta)}(x_2^n)^{-(I-\eta)}.$$

But, $(I-\eta)>0$ and, hence, the previous inequality violates Lipschitz continuity of $\alpha(x)$. This contradiction completes the proof of the theorem. □

5. Homogeneous systems. Consider the system

$$\dot{x}_1=f(x_1,x_2),\qquad \dot{x}_2=u, \tag{5.1}$$

where f is a homogeneous C^1-function of degree λ, i.e., $f(sx_1,sx_2)=s^\lambda f(x_1,x_2)$ for all real s. In order for various definitions to make sense, assume that λ is a rational number of the form $q/(2p+1)$ where p and q are nonnegative integers and $q\geqq 2p+1$.

Methods for determining the stability of two-dimensional homogeneous systems have been known for over 30 years. Consider the system

$$\dot{x}_1=\varphi_1(x_1,x_2),\qquad \dot{x}_2=\varphi_2(x_1,x_2), \tag{5.2}$$

where φ_1 and φ_2 are smooth homogeneous functions of degree $\lambda=q/(2p+1)$ $(q\geqq 2p+1)$.

THEOREM 5.1 [15], [16]. *The origin is a (globally) asymptotically stable equilibrium point of* (5.2) *if and only if q is odd and one of the following conditions is satisfied*:

(i) *System* (5.2) *does not have any one-dimensional invariant subspaces and*

$$\int_0^{2\pi}\frac{\cos\theta\phi_1(\cos\theta,\sin\theta)+\sin\theta\phi_2(\cos\theta,\sin\theta)}{\cos\theta\phi_2(\cos\theta,\sin\theta)-\sin\theta\phi_1(\cos\theta,\sin\theta)}\,d\theta<0$$

or

(ii) *System* (5.2) *is asymptotically stable on each of its one-dimensional invariant subspaces.*

The reader is referred to Hahn [15] for a very readable proof.

Remark 5.1. In the special case where φ_1 and φ_2 are homogeneous polynomials, Theorem 5.1 was proved by Haimo [16] using arguments which are more algebraic in nature than those found in Hahn [15].

Now the necessary and sufficient conditions for the asymptotic stabilizability of Theorem 5.1 can be stated. This result follows from Theorem 4.1 if f is C^w, but we give a slightly simpler proof in the general case and prove the global stabilizability.

THEOREM 5.2. *System* (5.1) *is locally asymptotically stabilizable by continuous feedback if and only if one of the following conditions hold*:

(i) *q is odd and there exist some $a \in \mathbb{R}^2_+$ such that $f(a) < 0$.*

(ii) *q is even and there exist points $a, b \in \mathbb{R}^2_+$ such that $f(a)f(b) < 0$.*

In both of these cases there exist globally asymptotically C^w stabilizing feedback.

Proof. Conditions given in the theorem are clearly necessary, for otherwise, either $\dot{x}_2(t) \geqq 0$ on $\mathbb{R}^2_+$ or $\dot{x}_2(t) \leqq 0$ on $\mathbb{R}^2_-$ along trajectories; therefore, the system is unstable at the origin regardless of feedback.

Sufficiency of (i). By using weak feedback equivalence with an associated coordinate transformation $(x_1, x_2) \mapsto (x_1, x_2 - \theta x_1)$ where θ is a real constant, it may be assumed that $f(1, 0) < 0$. Now let

$$u = -x_2^\lambda.$$

Now the system

$$\dot{x}_1 = f(x_1, x_2), \qquad \dot{x}_2 = -x_2^\lambda \tag{5.3}$$

obviously satisfies (ii) of Theorem 5.1 and this proves the sufficiency of (i) of Theorem 5.2.

Sufficiency of (ii). Now there exists $p = (p_1, p_2) \in \mathbb{R}^2_+$ such that $f(p) = 0$ and that $f(p_1 + \varepsilon, p_2)f(p_1 - \varepsilon, p_2) < 0$ for small positive ε. (Here we assumed (after applying a weak feedback transformation if necessary) that $p_2 \neq 0$.) Now a C^∞-function $\varphi : \mathbb{R} \to \mathbb{R}$ can be constructed such that $d\varphi/ds(0) = p_2/p_1$ and $x_1 f(x_1, \varphi(x_1)) < 0$ for all $x_1 \neq 0$.

Special case: f is C^2. In this case, the sufficiency of (ii) can be proved rather easily using the following ingenious "trick" found in Byrnes and Isidori [8] and in [23], [28], and [24].

Let us change coordinates as

$$y_1 = x_1, \qquad y_2 = x_2 - \varphi(x_1). \tag{5.4}$$

Now (5.1) is written in new coordinates as

$$\dot{y}_1 = f(y_1, y_2 + \varphi(y_1)), \qquad \dot{y}_2 = -\varphi'(y_1) f(y_1, y_2 + \varphi(y_1)) + u. \tag{5.5}$$

Let us write

$$f(y_1, y_2 + \varphi(y_1)) = f(y_1, \varphi(y_1)) + y_2 g(y_1, y_2),$$

where

$$g(y_1, y_2) = \int_0^1 \frac{\partial}{\partial y_2} f(y_1, ty_2 + \varphi(y_1))\, dt.$$

Since f is C^2, it follows that g is C^1.

Now define the feedback function for (5.5),

$$u = \varphi'(y_1)f(y_1, y_2+\varphi(y_1)) - y_2 - y_1 g(y_1, y_2).$$

It is now obvious that $V(y_1, y_2) = \frac{1}{2}(y_1^2+y_2^2)$ is a Lyapunov function for (3.5), proving the global asymptotic stabilizability of (5.1).

General case: f is C^1. The proof given above fails since g may now be merely continuous. However, the existence of globally asymptotically stabilizing C^∞-feedback in this general case can be proven by using a somewhat more complicated argument. (A similar construction is possible when (i) of Theorem 5.1 is satisfied as well.)

Assume without loss of generality that $p_1 = 1$. By replacing x_2 with $-x_2$ and u with $-u$, if necessary, it may be assumed that there exists some $\delta > 0$ such that

$$kf(1, k+p_2) < 0 \quad \text{for all } 0 < |k| \leqq \delta.$$

Without loss of generality, the assumption can be made that $|\varphi(x_1)| \leqq (\delta/2)|x_1|$ for all $x_1 \in \mathbb{R}$.

Let n be an odd integer greater than $2p/(2q+1)$. We will now prove that the feedback,

$$u = (\varphi(x_1) - x_2) + (\varphi(x_1) - x_2)^n,$$

globally asymptotically stabilizes the system (5.1).

Local asymptotic stability. In order to simplify the proof, consider the case $p_2 > 0$. The cases $p_2 = 0$ and $p_2 < 0$ can be handled by obvious modification of the proof and the details are left to the reader.

Without loss of generality assume that $\delta < p_2$. Let $\beta > 0$ and consider the region A^β, bounded by the lines

$$\begin{aligned}
\sigma_1^\beta &\equiv x_1 = \beta, \\
\sigma_2^\beta &\equiv x_2 - \beta(p_2+\delta) = p_2(x_1 - \beta), \\
\sigma_3^\beta &\equiv x_1 = -\beta, \\
\sigma_4^\beta &\equiv x_2 = -\beta p_2, \\
\sigma_5^\beta &\equiv x_2 = -\beta p_2 + 2p_2 x_1.
\end{aligned}$$

It may be claimed that when β is small enough, A^β is a positively invariant set. Use μ_i^β to denote the portion of the boundary of A^β which lies along σ_i^β. Clearly, the vector field $F := [f(x), (\varphi(x_1) - x_2) + (\varphi(x_1) - x_2)^n]^T$ points into A^β on μ_1^β, μ_3^β, and μ_4^β for all $\beta > 0$. Since $\varphi(x_1) - x_2 < 0$ on μ_2^β and $\varphi(x_1) - x_2 > 0$ on μ_5^β, it now suffices to prove that

$$\left| \frac{f(x)}{\varphi(x_1) - x_2} \right| \leqq \frac{1}{2p_2}$$

on μ_2^β and μ_5^β. Since $f(x)/(\varphi(x_1) - x_2) = 0$ at $(\beta, p_2\beta)$ and at $(-\beta, -(p_2-\delta)\beta)$, consider a point x in

$$(\mu_2^\beta \backslash (-\beta, -(p_2-\delta)\beta)) \cup (\mu_5^\beta \backslash (\beta, p_2\beta)),$$

and write $x = (r\cos\theta, r\sin\theta)$. Define θ_0 by $\cos\theta_0 = 1/\sqrt{1+p_2^2}$ and $\sin\theta_0 = p_2/\sqrt{1+p_2^2}$. Then

$$\begin{aligned}
\left| \frac{f(x)}{\varphi(x_1) - x_2} \right| &\leqq 2 \left| \frac{f(r\cos\theta, r\sin\theta)}{x_2 - p_2 x_1} \right| \\
&= \frac{2r^{2p/2q+1}|\gamma(\theta)|}{r\sqrt{(1+p_2^2)}\,|\sin(\theta - \theta_0)|},
\end{aligned}$$

where $\gamma(\theta)=f(\cos\theta,\sin\theta)$. Since f is C^1 it follows that γ is $C^1$4 and also $\gamma(\theta_0)=0$. Hence $|\gamma(0)|\leqq c|(\theta-\theta_0)|$, where c is a constant. Moreover, in the region under consideration $0<|\theta-\theta_0|<\pi-\nu$, where $\nu=\tan^{-1}(\delta)$, we can find a constant $d>0$ such that $|\sin(\theta-\theta_0)|\geqq d|(\theta-\theta_0)|$. Therefore,

$$\left|\frac{f(x)}{\varphi(x_1)-x_2}\right|\leqq\frac{c}{2d}\,\frac{r^{(2p/(2q+1)-1)}}{\sqrt{1+p_2^2}}.$$

Since $2p/2q+1>1$, it follows that when r is small enough

$$\frac{f(x)}{\varphi(x_1-x_2)}\leqq\frac{1}{2p_2},$$

if $x\in\mu_2^\beta\cup\mu_5^\beta$. Now, when β is small enough the required bound on r can be achieved. This concludes the proof that when β is small enough, A^β is a positively invariant set.

Now A^β is a compact positively invariant set. The origin is the only equilibrium point in A^β. It is clear that (since $|\varphi(x_1)|\leqq(\delta/2)x_1$) the region in A^β bounded by the straight lines $x_1=\beta$, $x_2=p_2x_1$, and $x_2=(p_2+\delta)x_1$ is an invariant set for a possibly smaller β. This precludes the existence of periodic orbits in A^β; therefore, we conclude that $w(x^0)$, the positive limit set of x^0, is the origin for all $x^0\in A^\beta$. Along with the positive invariance of A^β, this proves the local asymptotic stability of our system.

Global asymptotic stability. It must be shown that for all $x^0\in\mathbb{R}^2$, $w(x^0)=0$. By arguing as before but using the term $(\varphi(x_1)-x_2)^n$ in the feedback function as the dominant term whenever $(\varphi(x_1)-x_2)^n>1$, we prove that A^β is a positively invariant set for large values of β. The fact that the feedback function $(\varphi(x_1)-x_2)+(\varphi(x_1)-x_2)^n$ is positive on the line $x_2=p_2x_1$ precludes the existence of periodic orbits in A^β which enclose the origin. Therefore, for all $x^0\in A^\beta$ (β is large enough), $w(x^0)=0$. This completes the proof of Theorem 5.1. □

Remark 5.2. It is clear that

$$\varphi(x_1)=p_2x_1+\frac{\delta}{(\pi)^2}(\tan^{-1}(x_1))^2$$

satisfies the requirements of φ in the proof of Theorem 5.1, thereby proving the existence of real analytic feedback which stabilizes the system. Furthermore,

$$\varphi(x_1)=p_2x_1+x_1^2$$

yields polynomial feedback which locally asymptotically stabilizes the system. If f is homogeneous of odd degree, then the proof given above can be modified to show that the linear feedback $u=\gamma((p_2+\delta/2)x_1-x_2)$ (δ is a positive constant which is large enough) locally asymptotically stabilizes the system and polynomial feedback $u=\delta((p_2+\delta/2)x_1-x_2+((p_2+\delta/2)x_1-x_2)^n)$, where n is an odd integer greater than $2p/(2q+1)$, globally asymptotically stabilizes the system. (Here it is assumed that $kf(1,(k+p_2))<0$ for $0<|k|<\delta$. If $kf(1,(k+p_2))>0$ for $0<|k|<\delta$, then it is necessary to replace δ with $-\delta$ in the expression for feedback.)

Remark 5.3. In a recent paper Andreini, Bacciotti, and Stefani [2] gave sufficient conditions for stabilizability of homogeneous polynomial systems of odd degree in arbitrary dimensions. In the two-dimensional case their conditions are sufficient as well and they are equivalent to (i) of the previous theorem.

6. Local and global asymptotic stabilization of a special class of polynomial systems. Consider a system of the form

$$\dot{x}_1=ax_1^n+bx_2^m,\qquad \dot{x}_2=u, \tag{6.1}$$

where n and m are integers and a and b are real numbers. This special class is considered here since we can make global statements regarding this class.

Let $f(x) = ax_1^n + bx_2^m$. By Theorem 3.1, the following condition is necessary and sufficient for continuous feedback stabilization.

(*) There exists $q^1 \in \mathbb{R}^2_+$ such that $f(q^1) < 0$ and $q^2 \in \mathbb{R}^2_-$ such that $f(q^2) > 0$.

This is referred to as condition (*).

The objective in this section is to prove that condition (*) is also sufficient for global asymptotic stabilizability of (6.1). In the special case where $a > 0$, $b \neq 0$, $n = 1$, $m > 1$ and odd, the linearized system will have an eigenvalue in the open left half plane and, therefore, it follows easily that no C^1-feedback can stabilize the system. To avoid such cases it may be assumed in this section that $n > 1$ and $m > 1$. In most cases it will be shown that the stabilizing feedback can be constructed from the class of polynomials.

This class already displays interesting examples. The system

$$\dot{x}_1 = x_1 - x_2^3, \qquad \dot{x}_2 = u$$

is globally asymptotically stabilized using Hölder continuous feedback with Hölder exponent equal to $\frac{1}{3}$. But this is the maximum Hölder exponent. In particular, no Lipschitz continuous feedback stabilizes the system. As another example the system

$$\dot{x}_1 = x_1^2 - x_2^4, \qquad \dot{x}_2 = u$$

is C-globally asymptotically stabilizable but no C^3-feedback stabilizes it. We refer the reader to [12] for details on these examples.

We now state the main theorem of this section.

THEOREM 6.1. *Consider the system* (6.1) *and assume that* $n \geqq 2$ *and that condition* (*) *is satisfied. Then there exists* C^1 *globally asymptotically stabilizing feedback. If* $b \neq 0$ *and* m *is odd, then there exists globally asymptotically stabilizing polynomial feedback.*

Proof. We will prove the theorem by considering several different cases which encompass all of the possibilities.

Case 1. $a = 0$ or $b = 0$ or $n = m$. In this case $f(x)$ is a homogeneous polynomial of order m or n and Theorem 2.1 shows that the system is globally asymptotically stabilizable with C^∞ feedback. If $a \neq 0 \neq b$, $n = m$ and odd, Remark 5.2 points out that there exists globally asymptotically stabilizing polynomial feedback.

Case 2. $a \neq 0 \neq b$ and m is odd. The argument here is due to Kawski [18]. It is obvious that our system is weakly feedback equivalent to

$$\dot{x}_1 = \pm x_1^n - x_2^m, \qquad \dot{x}_2 = u. \tag{6.2}$$

Global asymptotic stabilizability of (6.2) can be proved rather easily by considering the Lyapunov function

$$V(x) = 4x_1^2 + \frac{1}{m+1} x_2^{m+1} \mp x_1^n x_2 + 4x_1^k, \tag{6.3}$$

and feedback

$$u = 8x_1 \mp n x_1^{n-1} x_2 + (\pm x_1^n - x_2^m) + 4k x_1^{k-1},$$

where k is an even integer which is greater than $n(m+1)/m$.

By using Hölders' inequality we obtain

$$|x_1^n x_2| = \left(\left(\frac{1}{2}x_2\right)^{m+1}\right)^{1/(m+1)} ((|2^{1/n}x_1|^{n(m+1)/m}))^{m/(m+1)}$$

$$\leqq \frac{1}{(m+1)}\left(\frac{1}{2}x_2\right)^{m+1} + \frac{m}{m+1}|x_1|^{n(m+1)/m}2^{(m+1)/m}.$$

If $|x_1| \leqq 1$, then

$$2^{(m+1)/m}\frac{m}{m+1}|x_1|^{n(m+1)/m} < 4|x_1|^2.$$

If $|x_1| > 1$, then

$$2^{(m+1)/m}\frac{m}{m+1}|x_1|\frac{n(m+1)}{m} < 4|x_1|^k. \tag{6.4}$$

Therefore, $V(x)$ is a positive-definite function on $\mathbb{R}^2$ which is radially unbounded. Moreover,

$$\dot{V}(x) = -(x_2^m \mp x_1^n)^2 \leqq 0.$$

Since the set $\{(x_1, x_2) \mid x_2^m \mp x_1^n = 0\}$ does not contain a nontrivial invariant set, it follows from the well-known LaSalle's theorem [17] that the system is globally asymptotically stable.

Case 3. $a \neq 0 \neq b$, n, m are even, $n > m$. It is obvious that system (6.1) is weakly feedback equivalent to

$$\dot{x}_1 = x_1^n - x_2^m, \qquad \dot{x}_2 = u. \tag{6.5}$$

We are going to prove the existence of C^1 feedback which globally asymptotically stabilizes (6.5).

Define $\psi : \mathbb{R} \to \mathbb{R}$ such that $s^n - (\psi(s))^m = 0$ and $s\psi(s) \geqq 0$ for $s \in \mathbb{R}$. Since $n > m$, it follows that ψ can be constructed to be a C^1-function. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a C^1-function such that:

(i) $\varphi(0) = 0$;

(ii) $0 < \varphi(s) < \frac{1}{2}|\psi(s)|$ for all $s \neq 0$ and $\varphi(s) > \frac{1}{4}|\psi(s)|$ for $|s| > 1$;

(iii) $\varphi(s) = \varphi(-s)$ and φ is monotone increasing for $s \geqq 0$.

Let k be an odd integer greater than n.

We claim that the feedback

$$u = \alpha(x) = \psi(x_1) + \varphi(x_1) - x_2 + (\psi(x_1) + \varphi(x_1) - x_2)^k$$

globally asymptotically stabilizes the system.

Local asymptotic stability. The idea is similar to the proof of Theorem 3.1. It is necessary to first produce arbitrarily small positively invariant sets enclosing the origin and then show that they cannot contain nontrivial periodic orbits. The shape of the invariant sets (called A^β) are depicted in Fig. 6.1 and the notation there is used to describe the boundaries of A^β.

Assume that β is small and positive and consider A^β as shown above. Now denote the vector field $[x_1^n - x_2^m, \alpha(x)]^T$ by $F(x)$. Obviously, F points into A^β on σ_1^β, σ_3^β, σ_4^β, σ_5^β, and σ_6^β. The problem is to construct σ_2^β in such a way that F points into A^β on σ_2^β and A^β is contained in a ball of radius $\delta(\beta)$ such that $\delta(\beta) \to 0$ as $\beta \to 0$. Consider the region in $\mathbb{R}_+^2$ between the two branches of the curve $x_1^n = x_2^m$ and inside a small

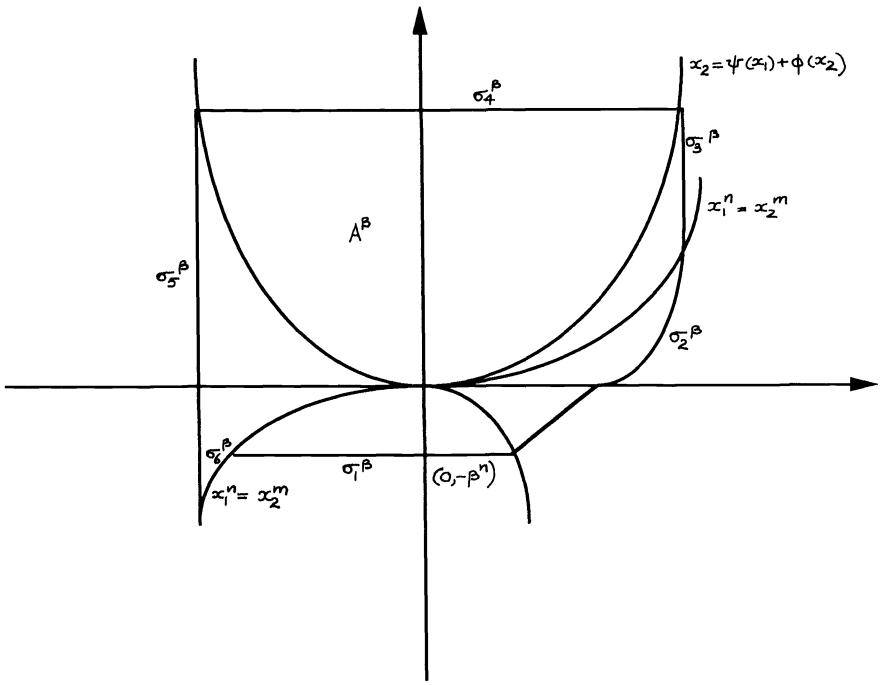


FIG. 6.1

ball. In this region

$$\left|\frac{x_1^n - x_2^m}{\alpha(x)}\right| \leqq \frac{|(\psi(x_1))^m - x_2^m|}{|\psi(x_1) - x_2|}$$

$$= \frac{|\psi(x_1) - x_2|}{|\psi(x_1) - x_2|} |\psi(x_1)^{m-1} + \psi(x_1)^{m-2} x_2 + \cdots + x_2^{m-1}|$$

$$< 1$$

when the radius of the ball is small enough. Therefore, it is possible to take σ_2^β to be the straight line with slope 1 through $(\beta^m, -\beta^n)$.

Now the positive invariance of A^β is established. Moreover, it is clear that A^β is contained in a ball of radius $\delta(\beta)$ where $\delta(\beta) \to 0$ as $\beta \to 0$. Since by construction, A^β does not contain equilibrium points other than at the origin (in order to prove the locally asymptotic stability) it now suffices to prove that there are no periodic orbits in A^β enclosing the origin. But this is obvious now since the region in $\mathbb{R}^2_-$ enclosed by σ_5^β and the branches of the curves $x_1^n = x_2^m$ is a positively invariant set.

Global asymptotic stability. Positive invariance of A^β is proved for large β by using the bound $\alpha(x) \geqq (\psi(x_1) - x_2 + \frac{1}{4}\psi(x_1))^k$, in a way similar to the above. Nonexistence of nontrivial periodic orbits follows by exactly the same reason as above.

Case 4. $a \neq 0 \neq b$, n, m even, $n < m$. Clearly, (6.1) is weakly feedback equivalent to

$$\dot{x}_1 = x_1^n - x_2^m, \qquad \dot{x}_2 = u. \tag{6.6}$$

The objective here is to prove the existence of globally asymptotically stabilizing C^1-feedback. It is shown in [12] that the case when $n=2$ and $m=4$ is not C^3-locally asymptotically stabilizable.

Let $\psi:\mathbb{R}\to\mathbb{R}$ be the C^1-function which satisfies the following properties:

(i) $\psi(s)^n - s^m = 0$ for all $s\in\mathbb{R}$

(ii) $s\psi(s) > 0$ for all $s\neq 0$.

Let $\varphi:\mathbb{R}\to\mathbb{R}$ be a C^1-function such that:

(i) $\varphi(0)=0$ and $\varphi(s)=+\varphi(-s)$ for all s

(ii) $0<\varphi(s)<\frac{1}{2}|\psi(s)|$ for all $s\neq 0$ and $\frac{1}{4}|\psi(s)|\leqq\varphi(s)$ for $|s|>1$.

It is claimed that the feedback function

$$u=\alpha(x)=x_1-\varphi(x_2)-\psi(x_2)+((x_1)-\varphi(x_2)-\psi(x_2))^{2m+1},$$

globally asymptotically stabilizes the system in this case.

The proof here is similar to that of Case 3. A^β turns out to be more complicated now, which is depicted by Fig. 6.2. We start with a small positive real $\beta>0$ and draw the boundaries of A^β such that σ_1^β and σ_2^β pass through $(\beta^m, -\beta^n)$. The boundary component σ_2^β will be described later.

Let $F(x)$ denote the vector field $[x_1^n - x_2^m, \alpha(x)]^T$. It is obvious that F points into A^β on all of the boundary components of A^β except possibly on σ_2^β. The aim is to construct σ_2^β in such a way that F points into A^β on σ_2^β and that the point of intersection of σ_2^β with the curve $x_1=\psi(x_2)$ approaches the origin as $\beta\to 0$. The difficulty here is that σ_2^β cannot be taken to be a straight line, for then it will not satisfy the second desired property of σ_2^β. However, in the region in $\mathbb{R}^2_+$ bounded by $x_1^n=x_2^m$ and inside

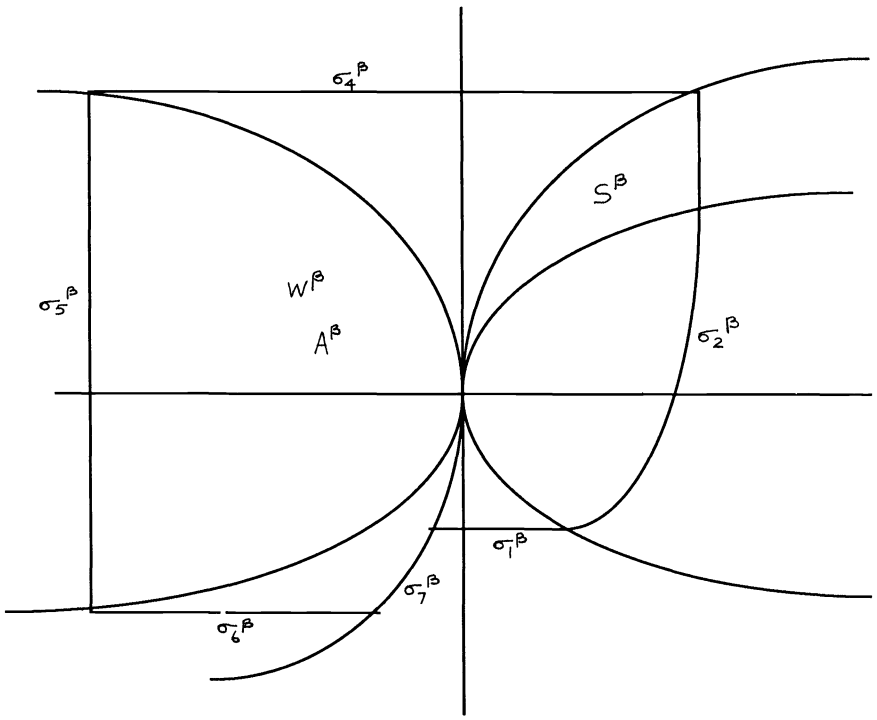


FIG. 6.2

a small ball

$$\left|\frac{x_1^n - x_2^m}{\alpha(x)}\right| \leqq \frac{x_1^n - x_2^m}{|x_1 - \psi(x_2)|}$$
$$= |x_1^{n-1} + x_1^{n-2}\psi(x_2) + \cdots + \psi(x_2)^{n-1}|$$
$$\leqq n x_1^{n-1} \quad (\text{since } x_1 \geqq \psi(x_2)).$$

Therefore, we take σ_2^β to be the solution of the differential equation

$$\frac{dx_2}{dx_1} = \frac{1}{nx_1^{n-1}}, \qquad x_1 \geqq \beta^m. \tag{6.7}$$

This analysis will be carried out under the assumption that $n > 2$. The case $n = 2$ is similar and for the sake of brevity that case is omitted. By solving (6.7) we get

$$x_2 = -\frac{1}{n(n-2)x_1^{n-2}} + \frac{1}{n(n-2)\beta^{m(n-2)}} - \beta^n, \qquad x_1 \geqq \beta^m. \tag{6.8}$$

Obviously, x_2 is monotone increasing from $-\beta^n$ to $(1/n(n-2)\beta^{m(n-2)}) - \beta^n$ as x_1 increases from β^m to ∞. It is claimed that when β is small enough this curve meets the $x_1 = \psi(x_2)$, $x_2 \geqq 0$ at some point $(\delta_1(\beta), \delta_2(\beta))$ and $(\delta_2(\beta), \delta_1(\beta)) \to 0$ as $\beta \to 0$.

Consider the function

$$\delta(x_2) = x_2 + \frac{1}{n(n-2)x_2^{(m(n-2)/n)}} - \frac{1}{n(n-2)\beta^{m(n-2)}} + \beta^n, \qquad x_2 > 0$$

and β is a very small positive constant. It now suffices to show that $\delta(\beta^n) > 0$ and $\delta(2^n\beta^n) < 0$. Now $\delta(\beta^n) = 2\beta^n > 0$ and

$$\delta(2^n\beta^n) = (2^n + 1)\beta^n - \frac{1}{n(n-2)\beta^{m(n-2)}}\left(1 - \frac{1}{2^{m(n-2)}}\right).$$

Since $n > 2$ it is now obvious that $\gamma(2^n\beta^n) < 0$ for all small enough β. This shows that the curve given by (6.8) and $x_1 = \psi(x_2)$, $x_2 \geqq 0$ meet at a point $(\delta_1, (\beta), \delta_2(\beta))$ where $\delta_2(\beta) \in (\beta^n, 2^n\beta^n)$ for small β and clearly $(\delta_1(\beta), \delta_2(\beta)) \to 0$ as $\beta \to 0$.

A positively invariant set A^β enclosing the origin is now produced which is contained in a ball of radius $r(\beta)$ where $r(\beta) \to 0$ as $\beta \to 0$. Furthermore, since the origin is the only equilibrium point in A^β and since $\alpha > 0$ on $x_1 = \psi(x_2)$ (which precludes the existence of periodic orbits enclosing the origin) it is concluded that the system is locally asymptotically stable. This concludes the proof of locally asymptotic stability of (6.1) with feedback $u = \alpha(x)$.

Global asymptotic stability is proved by showing that A^β, for large β (where σ_3^β is now defined to be a suitable straight line) is a positively invariant set without any periodic orbits.

Since the cases considered above are the only possibilities which can satisfy condition (*), Theorem 6.1 is now proven. □

7. Concluding remarks. Necessary and sufficient conditions are given for the asymptotic stabilizability for a class of two-dimensional C^∞-systems, which includes all real analytic systems. We also identify some obstructions for the existence of C^∞-stabilizing feedback in terms of some inequalities involving three numbers associated to the system. Currently, it is anticipated that these numbers have deeper geometric significance than the way in which they have been defined and may even give some related conditions for higher dimensional systems.

Acknowledgments. The authors gratefully acknowledge many fruitful discussions with C. I. Byrnes, M. Kawski, and E. Sontag during preparations for this paper. We also acknowledge the comments of an anonymous referee.

REFERENCES

[1] L. V. AHLFORS, *Complex Analysis*, Second edition, McGraw Hill, New York, 1966.
[2] A. ANDREINI, A. BACCIOTTI, AND G. STEFANI, *Global stabilizability of homogeneous vector fields of odd degree*, Systems Control Lett., 10 (1988), pp. 251-256.
[3] Z. ARTSTEIN, *Stabilization with relaxed controls*, Nonlinear Anal. Theory Methods Appl., 7 (1983), pp. 1163-1173.
[4] D. AYELS, *Stabilization of a class of nonlinear systems by a smooth feedback*, Systems Control Lett., 5 (1985), pp. 181-191.
[5] W. M. BOOTHBY AND R. MARINO, *Feedback stabilization of planar nonlinear systems*, Systems Control Lett., 12 (1989), pp. 87-92.
[6] R. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, Birkhauser, Boston, 1983.
[7] C. I. BYRNES AND A. ISIDORI, *Attitude stabilization of rigid spacecraft*, preprint.
[8] ———, *The analysis and design of nonlinear feedback systems* I, II: *Zero dynamics and global normal forms*, preprint.
[9] ———, *A frequency domain philosophy for nonlinear systems*, Proc. 23rd Annual IEEE Conference on Decision and Control, Las Vegas, NV, IEEE Computer Society, Washington, DC, 1984, pp. 1569-1573.
[10] J. CARR, *Applications of Center Manifold Theory*, Springer-Verlag, New York, 1981.
[11] W. P. DAYAWANSA AND C. F. MARTIN, *Asymptotic stabilization of two dimensional real analytic systems*, Systems Control Lett., 12 (1989), pp. 205-211.
[12] ———, *Two examples of stabilizable second order systems*, in Proc. Montana Conference on Computation and Control, Montana State University, Bozeman, MT, June 1988.
[13] M. GOLUBITSKY AND D. G. SCHÄFFER, *Singularities and Groups in Bifurcation Theory*, Vol. 1, Springer-Verlag, New York, 1985.
[14] P. GRIFFITHS AND J. HARRIS, *Principles of Algebraic Geometry*, John Wiley, New York, 1978.
[15] W. HAHN, *Stability of Motion*, Springer-Verlag, New York, 1967.
[16] V. T. HAIMO, *An algebraic approach to nonlinear stabilization*, Nonlinear Anal. Theory Methods Appl., 10 (1986).
[17] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.
[18] M. KAWSKI, *Stabilization of nonlinear systems in the plane*, preprint.
[19] D. E. KODITSCHEK, *Adaptive techniques for mechanical systems*, in Proc. 5th Yale Workshop on Adaptive Systems, Yale University, New Haven, CT, 1987, pp. 259-265.
[20] M. A. KROSNOSEL'SKII AND P. P. ZABREIKO, *Geometric Methods of Nonlinear Analysis*, Springer-Verlag, New York, 1984.
[21] S. LEFSCHETZ, *Algebraic Geometry*, Princeton University Press, NJ, 1953.
[22] R. MARINO, *Feedback stabilization of single-input nonlinear systems*, Systems Control Lett., 10 (1988), pp. 201-206.
[23] J. N. MATHER, *Stability of C^∞-mappings*, I. II, Ann. of Math., 87 (1962), pp. 89-104; 89 (1969), pp. 254-291.
[24] E. D. SONTAG, *Further facts about input to state stabilization*, Report 88-15, SYCON—Rutgers Center for Systems and Control, Rutgers University, New Brunswick, NJ, December 1988.
[25] E. D. SONTAG AND H. J. SUSSMANN, *Remarks on continuous feedback*, in Proc. IEEE Conference on Decision and Control, Albuquerque, December 1980, IEEE Computer Society, Washington, DC, pp. 916-921.
[26] R. SU, *On the equivalence of nonlinear systems*, Systems Control Lett., 2 (1982), pp. 48-52.
[27] R. THOM AND H. LEVINE, *Singularities of Differentiable Mappings*, Lecture Notes in Mathematics, Vol. 192, Springer-Verlag, New York, 1971.
[28] J. TSINIAS, *Sufficient Lyapunov like conditions for stabilization*, Math. Control Signals Systems, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1350-1371, November 1990

005

A J-SPECTRAL FACTORIZATION APPROACH TO $\mathscr{H}_\infty$ CONTROL*

MICHAEL GREEN†, KEITH GLOVER‡, DAVID LIMEBEER†, AND JOHN DOYLE§

Abstract. Necessary and sufficient conditions for the existence of suboptimal solutions to the standard model matching problem associated with $\mathscr{H}_\infty$ control are derived using J-spectral factorization theory. The existence of solutions to the model matching problem is shown to be equivalent to the existence of solutions to two coupled J-spectral factorization problems, with the second factor providing a parametrization of all solutions to the model matching problem. The existence of the J-spectral factors is then shown to be equivalent to the existence of nonnegative definite, stabilizing solutions to two indefinite algebraic Riccati equations, allowing a state-space formula for a linear fractional representation of all controllers to be given. A virtue of the approach is that a very general class of problems may be tackled within a conceptually simple framework, and no additional auxiliary Riccati equations are required.

Key words. $\mathscr{H}_\infty$ control, J-spectral factorization, indefinite factorization, four block problems, Riccati equations, Nehari's Theorem

AMS(MOS) subject classifications. 93C35, 47A68

Introduction. Since their inception, $\mathscr{H}_\infty$ control problems have been amenable to a variety of solution techniques. These range from the complex function theory approaches based on Nevanlinna-Pick-Schur interpolation to operator theoretic and state space approaches to $\mathscr{L}_\infty$ extension problems. In the case of simple problems, like sensitivity minimization, the relationships between these various approaches are well understood [8], [10], [14], [18]. The considerable body of knowledge about $\mathscr{H}_\infty$ control problems and their solution has evolved from the interaction between these various approaches, all of which provide solutions to the simple "Nehari type" problems which are conceptually elegant and computationally tractable. Unfortunately, this class of problems is too special to be of general engineering significance. In the case of more general problems, such as the mixed sensitivity problem, the mathematical solution was until recently more complicated, the interconnections were not well understood, and the computational burden associated with the solution was all but prohibitive (see [8], [10], [20]).

The J-spectral factorization approach to the problem of finding all suboptimal controllers for the simple "Nehari type" problems is well documented [2], [4], [10] and the approach has also been used to solve the optimal case [3]. In a recent paper [1], a general class of $\mathscr{H}_\infty$ control problems is solved via several spectral and J-spectral factorizations. The resulting algorithm is far from computationally simple. The new solution to the $\mathscr{H}_\infty$ problem presented in [12], however, requires just two indefinite algebraic Riccati equations to be solved and it was observed that these were associated with two J-factorizations.

In this paper we re-analyze the work in [1], showing that all the spectral and J-spectral factorizations can be subsumed into just two J-spectral factorizations. The Bart, Gohberg, and Kaashoek factorization theory [6] can then be used to associate the existence of the appropriate J-spectral factors with the solvability of two indefinite algebraic Riccati equations, and these can then be used to construct a generator of all solutions.

* Received by the editors December 27, 1988; accepted for publication (in revised form) October 27, 1989.
† Department of Electrical Engineering, Imperial College, London SW7 2BT, United Kingdom.
‡ Engineering Department, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom.
§ Department of Electrical Engineering, California Institute of Technology, Pasadena, California 91125.

Concurrent with this work, several of the other approaches to $\mathscr{H}_\infty$ control have been generalized and entirely new connections have been uncovered. The following remarks, which are in no way a complete survey, are intended to connect this paper with these other developments.

The four block distance problem has been solved by Glover, Limebeer, Doyle, Kasenally, and Safonov [12], [13], [21] using all-pass embedding. In Glover and Doyle [12] the equivalence between maximum entropy $\mathscr{H}_\infty$ control and risk sensitive control was established, a connection observed also in [7]. Moreover, Doyle et al. [9] have developed a state-space approach with a separation argument reminiscent of classical linear quadratic Gaussian (LQG) theory. Khargonekar, Petersen, and Rotea [16] have also considered a state feedback approach, observing a connection with LQ game theory. The connection between game theory and J-spectral factorization is long standing [5]. Extensions to time-varying systems using the maximum principle [25] and LQ game theory [19] have also been made. A conjugation approach developed by Kimura [17] is related to the J-spectral factorization method pursued here.

Note, however, that the assumptions used in the various approaches above are not all equivalent. In particular, the assumptions used here are more general than [9], where stronger assumptions are used for expository reasons. The optimal case is considered only in [13], [21].

Section 1 contains preliminaries and the standard stabilizing controller parametrization theory. In § 2 we analyze model matching problems of Nehari, unilateral and bilateral type and solve these in turn via J-spectral factorization. In order to satisfy the stability requirements it is necessary to impose an additional hitherto "unnoticed" condition on the J-spectral factors. Specifically, we will require the (1, 1) block of the factors to be outer. We note that Petersen and Clements [22] have also recently and independently observed that a J-spectral factorization with outer (1, 1) block can be associated with an $\mathscr{H}_\infty$ state feedback problem.

The relationship between J-spectral factorization and indefinite algebraic Riccati equations is analyzed in § 3. The results are reminiscent of existing results relating spectral factorization and Riccati equations and are derived using canonical factorization theory [6]. These results provide a state-space solution of the model matching problem in § 4. Section 5 gives necessary and sufficient conditions for a solution to the $\mathscr{H}_\infty$ control problem to exist and a representation formula for all solutions.

1. Preliminaries.

1.1 Notation.

$\mathbb{R}, \mathbb{C}$	real and complex number fields
$\bar{s}$	complex conjugate of $s \in \mathbb{C}$
$\mathscr{R}$	proper rational functions of a complex variable with complex coefficients
$\mathbb{C}^{m\times n}, \mathscr{R}^{m\times n}$	$m \times n$ matrices with entries in $\mathbb{C}$, $\mathscr{R}$
A^*	complex conjugate transpose of $A \in \mathbb{C}^{m\times n}$
$\lambda_i(A)$	ith eigenvalue of $A \in \mathbb{C}^{n\times n}$
$\lambda_{\max}(A)$	largest eigenvalue of a matrix $A \in \mathbb{C}^{n\times n}$
$\mathrm{In}(A)$	inertia of $A \in \mathbb{C}^{n\times n}$: $\mathrm{In}(A) = (\pi(A), \nu(A), \delta(A))$ where $\pi(A)$, $\nu(A)$, and $\delta(A)$ are, respectively, the number of eigenvalues of A in the open right and left half planes and on the imaginary axis
$A \geqq B, A > B$	$A - B \in \mathbb{C}^{n\times n}$ symmetric and positive semidefinite, positive definite
$\mathbf{M} \geqq \mathbf{N}, \mathbf{M} > \mathbf{N}$	$\mathbf{M} - \mathbf{N} \in \mathscr{R}^{n\times n}$ and $\mathbf{M}(j\omega) \geqq \mathbf{N}(j\omega)$, $\mathbf{M}(j\omega) > \mathbf{N}(j\omega)$, $\forall \omega \in \mathbb{R} \cup \infty$

$\mathcal{RL}_\infty^{m\times n}$	matrices in $\mathcal{R}^{m\times n}$ without imaginary axis poles
$\|\mathbf{M}\|_\infty$	$\mathcal{RL}_\infty$ norm: for $\mathbf{M}\in\mathcal{RL}_\infty \|\mathbf{M}\|_\infty = \sup_\omega \{\lambda_{\max}[\mathbf{M}(j\omega)^*\mathbf{M}(j\omega)]\}^{1/2}$
$\mathcal{RH}_\infty^{m\times n}$	subspace of $\mathcal{RL}_\infty^{m\times n}$ matrices without poles in the right half plane
$\mathcal{GH}_\infty^n$:	units of $\mathcal{RH}_\infty^{n\times n}$: $\mathbf{M}\in\mathcal{GH}_\infty^n \Leftrightarrow \mathbf{M}, \mathbf{M}^{-1}\in\mathcal{RH}_\infty^{n\times n}$
$\mathbf{M}^\sim$	$\mathbf{M}^\sim(s)=\mathbf{M}(-\bar{s})^*$
$\Gamma_\mathbf{M}$	Hankel operator with symbol $\mathbf{M}\in\mathcal{RL}_\infty^{m\times n}$

Associated with a matrix $\mathbf{M}\in\mathcal{R}^{m\times n}$ is a state space realization:

$$\mathbf{M}(s)=D+C(sI-A)^{-1}B \overset{s}{=} \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \in \mathbb{C}^{(n+m)\times(n+p)}. \tag{1.1}$$

If $\mathbf{P}\in\mathcal{R}^{(l+m)\times(p+q)}$ is partitioned as

$$\mathbf{P}=\begin{array}{c} \begin{matrix} p & q \end{matrix} \\ \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{bmatrix} \end{array} \begin{matrix} l \\ m \end{matrix} \tag{1.2}$$

then

$$\mathcal{F}(\mathbf{P},\mathbf{K})=\mathbf{P}_{11}+\mathbf{P}_{12}\mathbf{K}(I-\mathbf{P}_{22}\mathbf{K})^{-1}\mathbf{P}_{21}.$$

We say $\mathbf{P}$ is stabilizable if there exists such a $\mathbf{K}$ for which $\mathcal{F}(\mathbf{P},\mathbf{K})$ is internally stable (see [10]). The $\mathcal{H}_\infty$ control problem we will be concerned with is to find necessary and sufficient conditions for the existence of an internally stabilizing controller $\mathbf{K}$ such that $\|\mathcal{F}(\mathbf{P},\mathbf{K})\|_\infty<\gamma$, and when such conditions hold, to parametrize all solutions.

Finally, define the indefinite matrix $J_{pq}(\gamma)\in\mathbb{C}^{p+q}$, $\gamma>0$, by

$$J_{pq}(\gamma)=\begin{bmatrix} I_p & 0 \\ 0 & -\gamma^2 I_q \end{bmatrix}. \tag{1.3}$$

For convenience we will often abbreviate $J_{pq}(\gamma)$ to J.

1.2. Parametrization of all stabilizing controllers and the model matching problem. Suppose $\mathbf{P}\in\mathcal{R}^{(l+m)\times(p+q)}$ is partitioned as in (1.2) and is stabilizable. Suppose $\mathbf{P}_{22}$ has a doubly coprime factorization over $\mathcal{RH}_\infty$:

$$\mathbf{P}_{22}=\mathbf{N}_r\mathbf{D}_r^{-1}=\mathbf{D}_l^{-1}\mathbf{N}_l \tag{1.4a}$$

where

$$\begin{bmatrix} \mathbf{V}_r & \mathbf{U}_r \\ -\mathbf{N}_l & \mathbf{D}_l \end{bmatrix}\begin{bmatrix} \mathbf{D}_r & -\mathbf{U}_l \\ \mathbf{N}_r & \mathbf{V}_l \end{bmatrix}=\begin{bmatrix} I_q & 0 \\ 0 & I_m \end{bmatrix} \tag{1.4b}$$

is the corresponding Bezout identity. Further

$$\begin{bmatrix} \mathbf{V}_r & \mathbf{U}_r \\ -\mathbf{N}_l & \mathbf{D}_l \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{D}_r & -\mathbf{U}_l \\ \mathbf{N}_r & \mathbf{V}_l \end{bmatrix} \in \mathcal{RH}_\infty^{m+q}. \tag{1.4c}$$

It is well known (see, e.g., [8], [10], [23]) that $\mathbf{K}$ is a stabilizing controller if and only if $\mathbf{K}$ is given by

$$\mathbf{K}=\mathbf{K}_1\mathbf{K}_2^{-1}, \begin{bmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \end{bmatrix}=\begin{bmatrix} \mathbf{D}_r & -\mathbf{U}_l \\ \mathbf{N}_r & \mathbf{V}_l \end{bmatrix}\begin{bmatrix} \mathbf{Q} \\ I_m \end{bmatrix} \quad \mathbf{Q}\in\mathcal{RH}_\infty^{q\times m}. \tag{1.5}$$

Substituting (1.4) and (1.5) into $\mathcal{F}(\mathbf{P},\mathbf{K})$ we obtain

$$\begin{aligned} \mathcal{F}(\mathbf{P},\mathbf{K}) &= \mathbf{P}_{11}+\mathbf{P}_{12}\mathbf{K}(I-\mathbf{P}_{22}\mathbf{K})^{-1}\mathbf{P}_{21} \\ &= (\mathbf{P}_{11}-\mathbf{P}_{12}\mathbf{U}_l\mathbf{D}_l\mathbf{P}_{21})+(\mathbf{P}_{12}\mathbf{D}_r)\mathbf{Q}(\mathbf{D}_l\mathbf{P}_{21}) \\ &= \mathbf{T}_{11}+\mathbf{T}_{12}\mathbf{Q}\mathbf{T}_{21}. \end{aligned} \tag{1.6}$$

Thus, the $\mathscr{H}_\infty$ control problem can be posed as a model matching problem: Given the $\mathbf{T}_{ij}$'s, find necessary and sufficient conditions for the existence of $\mathbf{Q} \in \mathscr{RH}_\infty$ such that $\|\mathbf{T}_{11}+\mathbf{T}_{12}\mathbf{Q}\mathbf{T}_{21}\|_\infty < \gamma$ and, when such conditions hold, parametrize all solutions.

2. Model matching theory. In this section we solve a sequence model matching problem of increasing generality via J-spectral factorization. The existence of a solution to the model matching problem is shown to be equivalent to the existence of a J-spectral factor $\mathbf{W} \in \mathscr{GH}_\infty$ satisfying a relation of the form $\mathbf{G}^\sim\mathbf{J}\mathbf{G} = \mathbf{W}^\sim\mathbf{J}\mathbf{W}$ in which $\mathbf{W}_{11} \in \mathscr{GH}_\infty$, where $\mathbf{W}_{11}$ is the (1, 1) block of $\mathbf{W}$. The J-spectral factor $\mathbf{W}$, when it exists, is shown to parametrize all solutions to the model matching problem.

2.1. The Nehari problem. The purpose of this section is to summarize the standard results [2], [10] relating the Nehari extension problem to J-spectral factorization. The condition $\mathbf{W}_{11} \in \mathscr{GH}_\infty$ is new, however, and is one that not only turns out to be particularly useful in the more general model matching problems we subsequently consider, but simplifies the proofs for the Nehari case as well.

THEOREM 2.1. *Let* $\mathbf{R} \in \mathscr{RL}_\infty^{p\times q}$. *The following are equivalent*:

1. $\|\Gamma_{\mathbf{R}}\| < \gamma$;
2. *There exists* $\mathbf{Q} \in \mathscr{RH}_\infty^{p\times q}$ *such that* $\|\mathbf{R}+\mathbf{Q}\|_\infty < \gamma$;
3. *There exists* $\mathbf{W} \in \mathscr{GH}_\infty^{p+q}$ *with* $\mathbf{W}_{11} \in \mathscr{GH}_\infty^{p}$ *satisfying*

$$\mathbf{G}^\sim J_{pq}(\gamma)\mathbf{G} = \mathbf{W}^\sim J_{pq}(\gamma)\mathbf{W}, \qquad \mathbf{G} = \begin{bmatrix} I_p & \mathbf{R} \\ 0 & I_q \end{bmatrix}. \tag{2.1}$$

Proof. 1⇔2 is Nehari's Theorem. We shall prove that 1⇒3 and that 3⇒2.

3⇒2: Suppose a $\mathbf{W}$ with the required properties exists. Let $\mathbf{V} = \mathbf{W}^{-1}$ and partition $\mathbf{V}$ and $\mathbf{W}$ conformably with $\mathbf{G}$. Since $\mathbf{V}_{22}^{-1} = \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}$ [15, p. 656] and $\mathbf{W}_{11} \in \mathscr{GH}_\infty^p$, it follows that $\mathbf{V}_{22} \in \mathscr{GH}_\infty^q$. Set $\mathbf{Q} = \mathbf{V}_{12}(\mathbf{V}_{22})^{-1} \in \mathscr{RH}_\infty$, giving

$$\begin{bmatrix} \mathbf{R}+\mathbf{Q} \\ I \end{bmatrix} = \begin{bmatrix} I & \mathbf{R} \\ 0 & I \end{bmatrix}\begin{bmatrix} \mathbf{Q} \\ I \end{bmatrix} = \mathbf{G}\mathbf{V}\begin{bmatrix} 0 \\ \mathbf{V}_{22}^{-1} \end{bmatrix}.$$

Hence

$$\begin{aligned}(\mathbf{R}+\mathbf{Q})^\sim(\mathbf{R}+\mathbf{Q}) - \gamma^2 I &= \begin{bmatrix} \mathbf{R}+\mathbf{Q} \\ I \end{bmatrix}^\sim J \begin{bmatrix} \mathbf{R}+\mathbf{Q} \\ I \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ \mathbf{V}_{22}^{-1} \end{bmatrix}^\sim \mathbf{V}^\sim\mathbf{G}^\sim J\mathbf{G}\mathbf{V}\begin{bmatrix} 0 \\ \mathbf{V}_{22}^{-1} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{V}_{22}^{-1} \end{bmatrix}^\sim J \begin{bmatrix} 0 \\ \mathbf{V}_{22}^{-1} \end{bmatrix} \quad \text{by (2.1)} \\ &= -\gamma^2(\mathbf{V}_{22}\mathbf{V}_{22}^\sim)^{-1} < 0.\end{aligned}$$

This implies 2.

1⇒3: Decompose $\mathbf{R}$ as $\mathbf{R} = \mathbf{R}_+ + \mathbf{R}_-$, with $\mathbf{R}_-^\sim \in \mathscr{RH}_\infty$ and strictly proper, $\mathbf{R}_+ \in \mathscr{RH}_\infty$. Suppose, following [10], that $\mathbf{R}_-$ has a minimal realization $\mathbf{R}_-(s) = C(sI-A)^{-1}B$ and P and Q satisfy the Lyapunov equations

$$AP + PA^* = BB^* \tag{2.2a}$$

$$QA + A^*Q = C^*C. \tag{2.2b}$$

Since $\|\Gamma_{\mathbf{R}}\| < \gamma$, $\lambda_{\max}(QP) < \gamma^2$. Define

$$N = (I - \gamma^{-2}QP)^{-1}. \tag{2.3}$$

Define $\mathbf{X}$ by

$$\mathbf{X} \overset{s}{=} \left[\begin{array}{c|cc} -A^* & C^* & -QB \\ \hline \gamma^{-2}CPN & I & 0 \\ \gamma^{-2}B^*N & 0 & I \end{array}\right].$$

It is readily verified (using the state transformation $\left[\begin{smallmatrix} \gamma^2 N & Q \\ PN & I \end{smallmatrix}\right]^{-1}$ on $G^{\sim}JG_-$) that

$$G^{\sim}JG_- = \mathbf{X}^{\sim}J\mathbf{X}, \qquad \mathbf{G}_- = \begin{bmatrix} I & \mathbf{R}_- \\ 0 & I \end{bmatrix}.$$

Since $-A^*$ is asymptotically stable, we see that $\mathbf{X} \in \mathscr{RH}_\infty$. It is also easy to verify using (2.2) and (2.3) that the "A" matrix of $\mathbf{X}^{-1} = -N^{-1}A^*N$, so $\mathbf{X} \in \mathscr{GH}_\infty$. The "$A$" matrix of $(\mathbf{X}_{11})^{-1}$ is given by

$$\hat{A} = -A^* - \gamma^{-2}C^*CPN.$$

Using (2.2) and (2.3) it is easy to establish that

$$\hat{A}N^{-1}P^{-1} + P^{-1}N^*\hat{A}^* = -\begin{bmatrix} \gamma^{-1}C^* & BP^{-1} \end{bmatrix}\begin{bmatrix} \gamma^{-1}C \\ B^*P^{-1} \end{bmatrix}$$

which shows, since $N^{-1}P^{-1} > 0$, that $\hat{A}$ is asymptotically stable, and consequently $\mathbf{X}_{11} \in \mathscr{GH}_\infty$, provided $(\hat{A}, [\gamma^{-1}C^*\ BP^{-1}])$ is controllable [11, Thm. 3.3]. The required controllability is easily seen from

$$[\hat{A} \quad \gamma^{-1}C^*] = [-A^* \quad C^*]\begin{bmatrix} I & 0 \\ -\gamma^{-2}CPN & \gamma^{-1}I \end{bmatrix}.$$

Finally, observe that

$$\mathbf{G} = \begin{bmatrix} I & \mathbf{R} \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & \mathbf{R}_- \\ 0 & I \end{bmatrix}\begin{bmatrix} I & \mathbf{R}_+ \\ 0 & I \end{bmatrix},$$

so $\mathbf{W}$ given by

$$\mathbf{W} = \mathbf{X}\begin{bmatrix} I & \mathbf{R}_+ \\ 0 & I \end{bmatrix}$$

has the required properties. □

Note that, provided γ is not in the spectrum of $\Gamma_{\mathbf{R}}$, the generalization to the AAK problem (where $\mathbf{Q}$ is allowed k poles in the right half plane) is simply that $\mathbf{W}_{11}^{-1}$ is allowed k poles in the right half plane.

Consider the factorization (2.1). As with spectral factorization, $\mathbf{W} \in \mathscr{GH}_\infty$ satisfying (2.1) is not unique, being determined only up to a J-unitary matrix (see the following lemma). Supposing that *one* of these solutions has the property $\mathbf{W}_{11} \in \mathscr{GH}_\infty$, it is important to establish whether or not *all* of the other possible solutions have this property as well. For unless the property $\mathbf{W}_{11} \in \mathscr{GH}_\infty$ is an all or none affair, Theorem 2.1 will be of little practical value, as one would have to look through the class of possible $\mathbf{W}$'s in search of one with the desired $\mathbf{W}_{11} \in \mathscr{GH}_\infty$ property. Fortunately, this is not necessary.

LEMMA 2.2. *Suppose* $\mathbf{W} \in \mathscr{GH}_\infty^{p+q}$. *Then*

1. $\mathbf{Y} \in \mathscr{GH}_\infty^{p+q}$ *satisfies* $\mathbf{Y}^{\sim}\mathbf{J}\mathbf{Y} = \mathbf{W}^{\sim}\mathbf{J}\mathbf{W}$ *if and only if* $\mathbf{Y} = \mathbf{A}\mathbf{W}$, *where* $\mathbf{A}$ *is a constant J-unitary matrix* (*i.e.*, $A^*JA = J$).

2. *If* $\mathbf{W}_{11}^{\sim}\mathbf{W}_{11}-\gamma^2\mathbf{W}_{21}^{\sim}\mathbf{W}_{21}\geqq 0$ *and* $\mathbf{Y}\in\mathcal{GH}_\infty^{p+q}$ *satisfies* $\mathbf{Y}^{\sim}\mathbf{J}\mathbf{Y}=\mathbf{W}^{\sim}\mathbf{J}\mathbf{W}$, *then* $\mathbf{Y}_{11}\in\mathcal{GH}_\infty^{p}$ *if and only if* $\mathbf{W}_{11}\in\mathcal{GH}_\infty^{p}$.

Proof. Suppose $\mathbf{Y}\in\mathcal{GH}_\infty$ satisfies $\mathbf{W}^{\sim}\mathbf{J}\mathbf{W}=\mathbf{Y}^{\sim}\mathbf{J}\mathbf{Y}$. Then

$$(\mathbf{Y}^{\sim})^{-1}\mathbf{W}^{\sim}\mathbf{J}=\mathbf{J}\mathbf{Y}\mathbf{W}^{-1}. \tag{2.4}$$

Since $\mathbf{Y}\mathbf{W}^{-1}\in\mathcal{GH}_\infty$, it follows that $\mathbf{Y}\mathbf{W}^{-1}=A$ is constant and is J-unitary by (2.4). The converse is obvious.

Observe that $\mathbf{W}_{11}^{\sim}\mathbf{W}_{11}-\gamma^2\mathbf{W}_{21}^{\sim}\mathbf{W}_{21}\geqq 0$ and $\mathbf{W}_{11}\in\mathcal{GH}_\infty\Rightarrow\mathbf{W}_{21}\mathbf{W}_{11}^{-1}\in\mathcal{RH}_\infty$ and $\|\mathbf{W}_{21}\mathbf{W}_{11}^{-1}\|_\infty\leqq\gamma^{-1}$. Also $A^*JA=J\Rightarrow A^{-1}=J^{-1}A^*J\Rightarrow AJ^{-1}A^*=J^{-1}$, the (1, 1) block of which is $A_{11}A_{11}^*-\gamma^{-2}A_{12}A_{12}^*=I$. Hence A_{11} is nonsingular and $\|A_{11}^{-1}A_{12}\|<\gamma$. Therefore $(I+A_{11}^{-1}A_{12}\mathbf{W}_{21}\mathbf{W}_{11}^{-1})\in\mathcal{GH}_\infty$ or, equivalently, $\mathbf{Y}_{11}=A_{11}\mathbf{W}_{11}+A_{12}\mathbf{W}_{21}\in\mathcal{GH}_\infty$. For the converse, interchange $\mathbf{Y}$ and $\mathbf{W}$ in the above argument. □

Note that if $\mathbf{G}^{\sim}J\mathbf{G}=\mathbf{W}^{\sim}J\mathbf{W}$ and $\mathbf{G}_{21}=0$ then the condition $\mathbf{W}_{11}^{\sim}\mathbf{W}_{11}-\gamma^2\mathbf{W}_{21}^{\sim}\mathbf{W}_{21}\geqq 0$ is satisfied. Thus, given *any* $\mathbf{W}\in\mathcal{GH}_\infty$ such that $\mathbf{G}^{\sim}J\mathbf{G}=\mathbf{W}^{\sim}J\mathbf{W}$ with $\mathbf{G}$ as in (2.1), the Nehari problem has a solution if and only if $\mathbf{W}_{11}\in\mathcal{GH}_\infty$. The point is that if $\mathbf{W}_{11}\notin\mathcal{GH}_\infty$, we do not have to worry about the possibility of some other solution $\mathbf{Y}\in\mathcal{GH}_\infty$ such that $\mathbf{G}^{\sim}J\mathbf{G}=\mathbf{Y}^{\sim}J\mathbf{Y}$ having the property $\mathbf{Y}_{11}\in\mathcal{GH}_\infty$.

The next result is also standard [2], [10] and provides a characterization of all solutions to suboptimal Nehari extension problems.

THEOREM 2.3. *Let* $\mathbf{R}\in\mathcal{RL}_\infty^{p\times q}$ *and suppose there exists* $\mathbf{W}\in\mathcal{GH}_\infty^{p+q}$ *with* $\mathbf{W}_{11}\in\mathcal{GH}_\infty^{p}$ *satisfying* (2.1), *i.e.,* $\mathbf{G}^{\sim}J\mathbf{G}=\mathbf{W}^{\sim}J\mathbf{W}$. *Then the set of all matrices* $\mathbf{Q}\in\mathcal{RH}_\infty^{p\times q}$ *such that* $\|\mathbf{R}+\mathbf{Q}\|_\infty\leqq\gamma$ *is given by*

$$\mathbf{Q}=\mathbf{Q}_1\mathbf{Q}_2^{-1},\begin{bmatrix}\mathbf{Q}_1\\ \mathbf{Q}_2\end{bmatrix}=\mathbf{W}^{-1}\begin{bmatrix}\mathbf{U}\\ I_q\end{bmatrix},\quad \mathbf{U}\in\mathcal{RH}_\infty^{p\times q}\ \textit{with}\ \|\mathbf{U}\|_\infty\leqq\gamma. \tag{2.5}$$

Proof. Let $\mathbf{V}=\mathbf{W}^{-1}$ and recall $\mathbf{V}_{22}\in\mathcal{GH}_\infty$. Suppose $\mathbf{U}\in\mathcal{RH}_\infty$, $\|\mathbf{U}\|_\infty\leqq\gamma$. To prove $\mathbf{Q}\in\mathcal{RH}_\infty$ we show that $\mathbf{Q}_2\in\mathcal{GH}_\infty$. By (2.1), $\mathbf{V}J^{-1}\mathbf{V}^{\sim}=\mathbf{G}^{-1}J^{-1}(\mathbf{G}^{-1})^{\sim}$, the 2, 2 block of which gives $\mathbf{V}_{21}\mathbf{V}_{21}^{\sim}-\gamma^{-2}\mathbf{V}_{22}\mathbf{V}_{22}^{\sim}=-\gamma^{-2}I$. Hence $\|\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\|_\infty<\gamma^{-1}$. It follows that $(\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{U}+I)\in\mathcal{GH}_\infty$ and hence $\mathbf{Q}_2=\mathbf{V}_{22}(\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{U}+I)\in\mathcal{GH}_\infty$, for all $\mathbf{U}\in\mathcal{RH}_\infty$ with $\|\mathbf{U}\|_\infty\leqq\gamma$. Also, with $\mathbf{Q}$ defined by (2.5) we have

$$\begin{aligned}(\mathbf{R}+\mathbf{Q})^{\sim}(\mathbf{R}+\mathbf{Q})-\gamma^2 I&=(\mathbf{Q}_2^{-1})^{\sim}\begin{bmatrix}\mathbf{U}\\ I\end{bmatrix}^{\sim}\mathbf{V}^{\sim}\mathbf{G}^{\sim}J\mathbf{G}\mathbf{V}\begin{bmatrix}\mathbf{U}\\ I\end{bmatrix}\mathbf{Q}_2^{-1}\\ &=(\mathbf{Q}_2^{-1})^{\sim}[\mathbf{U}^{\sim}\mathbf{U}-\gamma^2 I]\mathbf{Q}_2^{-1}\leqq 0.\end{aligned}$$

Conversely, suppose $\mathbf{Q}\in\mathcal{RH}_\infty$ is such that $\|\mathbf{R}+\mathbf{Q}\|_\infty\leqq\gamma$. Define

$$\begin{bmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{bmatrix}=\mathbf{W}\begin{bmatrix}\mathbf{Q}\\ I\end{bmatrix}=\mathbf{W}\mathbf{G}^{-1}\begin{bmatrix}\mathbf{R}+\mathbf{Q}\\ I\end{bmatrix}\in\mathcal{RH}_\infty.$$

Observe that $\mathbf{U}_1,\mathbf{U}_2\in\mathcal{RH}_\infty$ are right coprime and that

$$\mathbf{U}_1^{\sim}\mathbf{U}_1-\gamma^2\mathbf{U}_2^{\sim}\mathbf{U}_2=\begin{bmatrix}\mathbf{R}+\mathbf{Q}\\ I\end{bmatrix}^{\sim}J\begin{bmatrix}\mathbf{R}+\mathbf{Q}\\ I\end{bmatrix}\leqq 0.$$

It follows that $\mathbf{U}_2$ is invertible in $\mathcal{RL}_\infty$, and that $\mathbf{U}=\mathbf{U}_1\mathbf{U}_2^{-1}\in\mathcal{RL}_\infty$ with $\|\mathbf{U}\|_\infty\leqq\gamma$. Hence (2.5) holds, with $\mathbf{Q}_2=\mathbf{U}_2^{-1}$, and $\mathbf{Q}_1=\mathbf{Q}\mathbf{Q}_2$ and it remains to show that $\mathbf{U}\in\mathcal{RH}_\infty$. This we do by showing that $\mathbf{U}_2\in\mathcal{GH}_\infty$. To see this, observe that, since $\|\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{U}\|_\infty\leqq\|\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\|_\infty\ \|\mathbf{U}\|_\infty<1$, the winding number (around the origin) of $\det\{(\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{U}+I)(j\omega)\}$ is zero. Also $\mathbf{V}_{22}^{-1}=(\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{U}+I)\mathbf{U}_2\in\mathcal{GH}_\infty$. It follows that the winding number of $\det(\mathbf{U}_2(j\omega))$ is zero, giving $\mathbf{U}_2\in\mathcal{GH}_\infty$, since $\mathbf{U}_2\in\mathcal{RH}_\infty$. □

2.2. The unilateral model matching problem. In the last section we considered a factorization problem associated with the Nehari extension problem $\|\mathbf{R}+\mathbf{Q}\|_\infty < \gamma$. In this case the factorization problem is particularly easy because $\mathbf{G}$ is square and invertible in $\mathcal{RL}_\infty$, a fact used in the proof of Theorem 2.3. We now turn to the unilateral model matching problem where we seek $\mathbf{Q} \in \mathcal{RH}_\infty$ such that $\|\mathbf{A}+\mathbf{BQ}\|_\infty < \gamma$, where $\mathbf{B}$ is "tall" (i.e., has more rows than columns), and the relevant "$\mathbf{G}$" is now also "tall." A related theorem is given in [14, p. 58].

The "tall" J-spectral factorization problem is shown to be equivalent to two spectral factorization problems together with a "square" J-spectral factorization problem (i.e., one of Nehari type). The techniques are similar to those used elsewhere [8], [10], [20] to reduce "two-block" distance problems to Nehari problems, but here the interpretation is in terms of the existence of solutions to J-spectral factorization problems.

THEOREM 2.4. *Suppose*

$$\mathbf{G} = \begin{bmatrix} \mathbf{B} & \mathbf{A} \\ 0 & I_q \end{bmatrix} \in \mathcal{RL}_\infty^{(l+q)\times(p+q)}$$

has a left inverse in $\mathcal{RL}_\infty$. *The following are equivalent*:

1. *There exists a* $\mathbf{Q} \in \mathcal{RH}_\infty^{p\times q}$ *such that* $\|\mathbf{A}+\mathbf{BQ}\|_\infty < \gamma$;
2. *There exists a* $\mathbf{W} \in \mathcal{GH}_\infty^{p+q}$ *with* $\mathbf{W}_{11} \in \mathcal{GH}_\infty^{p}$ *satisfying*

$$\mathbf{G}^{\sim} J_{lq}(\gamma)\mathbf{G} = \mathbf{W}^{\sim} J_{pq}(\gamma)\mathbf{W}. \tag{2.6}$$

Furthermore, if such a $\mathbf{W}$ *exists, the set of all matrices* $\mathbf{Q} \in \mathcal{RH}_\infty$ *satisfying* $\|\mathbf{A}+\mathbf{BQ}\|_\infty \leqq \gamma$ *is given by*

$$\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2^{-1}, \quad \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix} = \mathbf{W}^{-1}\begin{bmatrix} \mathbf{U} \\ I_q \end{bmatrix}, \quad \mathbf{U} \in \mathcal{RH}_\infty^{p\times q} \text{ with } \|\mathbf{U}\|_\infty \leqq \gamma. \tag{2.7}$$

Proof. $\mathbf{G}$ left invertible in $\mathcal{RL}_\infty$ is equivalent to $\mathbf{B}$ full column rank on the imaginary axis, so there exists $\mathbf{B}_0 \in \mathcal{GH}_\infty$ such that $\mathbf{B}_0^{\sim}\mathbf{B}_0 = \mathbf{B}^{\sim}\mathbf{B}$. Reduce to the Nehari problem as follows:

Let $\mathbf{B}_i = \mathbf{B}\mathbf{B}_0^{-1}$ and note $\mathbf{B}_i^{\sim}\mathbf{B}_i = I$. Let $\mathbf{B}_\perp$ be such that $[\mathbf{B}_i\mathbf{B}_\perp]$ is all-pass. Then

$$\begin{aligned} \|\mathbf{A}+\mathbf{BQ}\|_\infty \gamma &\Leftrightarrow \|\mathbf{A}+[\mathbf{B}_i\mathbf{B}_\perp]\begin{bmatrix} \mathbf{B}_0\mathbf{Q} \\ 0 \end{bmatrix}\|_\infty < \gamma \\ &\Leftrightarrow \left\| \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_0\mathbf{Q} \\ 0 \end{bmatrix} \right\|_\infty < \gamma, \qquad \mathbf{R} = [\mathbf{B}_i\mathbf{B}_\perp]^{\sim}\mathbf{A} \\ &\Leftrightarrow \|\mathbf{R}_2\|_\infty < \gamma \text{ and } (\mathbf{R}_1+\mathbf{B}_0\mathbf{Q})^{\sim}(\mathbf{R}_1+\mathbf{B}_0\mathbf{Q}) + \mathbf{R}_2^{\sim}\mathbf{R}_2 < \gamma^2 I. \end{aligned}$$

Thus, there exists $\mathbf{Q} \in \mathcal{RH}_\infty$ such that $\|\mathbf{A}+\mathbf{BQ}\|_\infty < \gamma$ if and only if:

$$\exists \mathbf{N} \in \mathcal{GH}_\infty \text{ with } \gamma^2\mathbf{N}^{\sim}\mathbf{N} = \mathbf{\Phi} = \gamma^2 I - \mathbf{R}_2^{\sim}\mathbf{R}_2 = \gamma^2 I_q - \mathbf{A}^{\sim}[I - \mathbf{B}(\mathbf{B}^{\sim}\mathbf{B})^{-1}\mathbf{B}^{\sim}]\mathbf{A}; \tag{2.8a}$$

and

$$\exists \hat{\mathbf{Q}}(=\mathbf{B}_0\mathbf{Q}\mathbf{N}^{-1}) \in \mathcal{RH}_\infty \text{ such that } \|\mathbf{R}_1\mathbf{N}^{-1}+\hat{\mathbf{Q}}\|_\infty < \gamma. \tag{2.8b}$$

By Theorem 2.1, there exists $\hat{\mathbf{Q}} \in \mathcal{RH}_\infty$ such that

$$\|\mathbf{R}_1\mathbf{N}^{-1}+\hat{\mathbf{Q}}\|_\infty < \gamma \Leftrightarrow \exists \mathbf{X} \in \mathcal{GH}_\infty \quad \text{with } \mathbf{X}_{11} \in \mathcal{GH}_\infty$$

such that

$$\begin{bmatrix} I & 0 \\ (N^{-1})^{\sim}\mathbf{R}_1^{\sim} & I \end{bmatrix} J \begin{bmatrix} I & \mathbf{R}_1\mathbf{N}^{-1} \\ 0 & I \end{bmatrix} = \mathbf{X}^{\sim} J \mathbf{X}.$$

Note also that $\mathbf{R}_1 = (\mathbf{B}_0^{\sim})^{-1}\mathbf{B}^{\sim}\mathbf{A}$. Now observe that

$$\mathbf{G}^{\sim} J \mathbf{G} = \begin{bmatrix} \mathbf{B}_0 & 0 \\ \mathbf{A}^{\sim}\mathbf{B}\mathbf{B}_0^{-1} & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -\mathbf{\Phi} \end{bmatrix} \begin{bmatrix} \mathbf{B}_0 & (\mathbf{B}_0^{\sim})^{-1}\mathbf{B}^{\sim}\mathbf{A} \\ 0 & I \end{bmatrix}. \tag{2.9}$$

It follows that $\mathbf{W}$ exists $\Leftrightarrow \mathbf{X}$ and $\mathbf{N}$ exists $(\mathbf{X} = \mathbf{W}[\begin{smallmatrix} \mathbf{B}_0 & 0 \\ 0 & \mathbf{N} \end{smallmatrix}]^{-1})$ and the theorem is proved.

That (2.7) gives all solutions now follows from Theorem 2.3. □

Remark 2.5. The condition that $\mathbf{G}$ (equivalently $\mathbf{B}$) has a left inverse in $\mathcal{RL}_\infty$ is not necessary for there to exist a solution to the model matching problem. It is, however, a necessary condition for the existence of $\mathbf{W} \in \mathcal{GH}_\infty$ such that $\mathbf{G}^\sim \mathbf{J}\mathbf{G} = \mathbf{W}^\sim \mathbf{J}\mathbf{W}$.

2.3. The bilateral model matching problem. We now extend the constructions of § 2.2 to the bilateral case. That is, we seek $\mathbf{Q} \in \mathcal{RH}_\infty$ such that $\|\mathbf{A}+\mathbf{BQC}\|_\infty < \gamma$, with $\mathbf{B}$ "tall" and $\mathbf{C}$ "wide." The technique is based on reduction to the unilateral case, and the result involves two J-spectral factorizations.

THEOREM 2.6. *Suppose* $\mathbf{A} \in \mathcal{RL}_\infty^{l\times p}$, $\mathbf{B} \in \mathcal{RL}_\infty^{l\times q}$ *and* $\mathbf{C} \in \mathcal{RL}_\infty^{m\times p}$. *Suppose also that* $\mathbf{B}$ *has a left inverse and* $\mathbf{C}$ *has a right inverse in the appropriate* $\mathcal{RL}_\infty$ *spaces. Let* $\mathbf{B} = \mathbf{B}_a\mathbf{B}_s$ *in which* $\mathbf{B}_a \in \mathcal{RL}_\infty^{l\times l}$ *is all-pass and* $\mathbf{B}_s \in \mathcal{RH}_\infty^{l\times q}$. *Then there exists a* $\mathbf{Q} \in \mathcal{RH}_\infty^{q\times m}$ *such that* $\|\mathbf{A}+\mathbf{BQC}\|_\infty < \gamma$ *if and only if*

1. *There exists a* $\mathbf{V} \in \mathcal{GH}_\infty^{m+l}$ *with* $\mathbf{V}_{11} \in \mathcal{GH}_\infty^m$ *satisfying*

$$\mathbf{H}J_{pl}(\gamma)\mathbf{H}^\sim = \mathbf{V}J_{ml}(\gamma)\mathbf{V}^\sim, \qquad \mathbf{H} = \begin{bmatrix} \mathbf{C} & 0 \\ \mathbf{B}_a^\sim \mathbf{A} & I_l \end{bmatrix} \tag{2.10}$$

and

2. *There exists a* $\mathbf{W} \in \mathcal{GH}_\infty^{q+m}$ *with* $\mathbf{W}_{11} \in \mathcal{GH}_\infty^q$ *satisfying*

$$\mathbf{G}^\sim J_{lm}(\gamma)\mathbf{G} = \mathbf{W}^\sim J_{qm}(\gamma)\mathbf{W}, \qquad \mathbf{G} = \hat{J}\mathbf{V}^{-1}\hat{J}^* \begin{bmatrix} \mathbf{B}_s & 0 \\ 0 & I_m \end{bmatrix} \tag{2.11}$$

where

$$\hat{J} = \begin{bmatrix} 0 & -I_l \\ I_m & 0 \end{bmatrix}. \tag{2.12}$$

In this case, the set of all matrices $\mathbf{Q} \in \mathcal{RH}_\infty^{q\times m}$ *such that* $\|\mathbf{A}+\mathbf{BQC}\|_\infty \leqq \gamma$ *is given by*

$$\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2^{-1}, \quad \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix} = \mathbf{W}^{-1}\begin{bmatrix} \mathbf{U} \\ I_m \end{bmatrix}, \quad \mathbf{U} \in \mathcal{RH}_\infty^q \text{ with } \|\mathbf{U}\|_\infty \leqq \gamma. \tag{2.13}$$

Proof. We may assume, without loss of generality, that $\mathbf{B} \in \mathcal{RH}_\infty$, since $\|\mathbf{A}+\mathbf{BQC}\|_\infty \leqq \gamma \Leftrightarrow \|\mathbf{B}_a^\sim \mathbf{A}+\mathbf{B}_s\mathbf{QC}\|_\infty \leqq \gamma$.

With $\mathbf{B} \in \mathcal{RH}_\infty$ we see that 1 is necessary by applying Theorem 2.4 to the problem $\mathbf{A}^* + \mathbf{C}^*\hat{\mathbf{Q}}$, where $\hat{\mathbf{Q}} = (\mathbf{BQ})^*$.

Let $\mathbf{C}_0 \in \mathcal{GH}_\infty$ be such that $\mathbf{CC}^\sim = \mathbf{C}_0\mathbf{C}_0^\sim$ and define $\mathbf{C}_i = \mathbf{C}_0^{-1}\mathbf{C}$. Let $\mathbf{C}_\perp$ be such that $\left[\begin{smallmatrix} \mathbf{C}_i \\ \mathbf{C}_\perp \end{smallmatrix}\right]$ is all-pass. Define $\mathbf{R}$ by

$$\mathbf{R} = [\mathbf{R}_1\mathbf{R}_2] = \mathbf{A}\begin{bmatrix} \mathbf{C}_i \\ \mathbf{C}_\perp \end{bmatrix}^\sim.$$

As in the proof of Theorem 2.4, the existence of $\mathbf{V}$ satisfying (2.10) implies that there exists $\mathbf{M} \in \mathcal{GH}_\infty$ such that

$$\gamma^2\mathbf{M}\mathbf{M}^\sim = \gamma^2 I - \mathbf{R}_2\mathbf{R}_2^\sim.$$

So $\mathbf{Q} \in \mathcal{RH}_\infty$ satisfies $\|\mathbf{A}+\mathbf{BQC}\|_\infty < \gamma \Leftrightarrow \mathbf{V}$ exists and $\|\mathbf{M}^{-1}\mathbf{R}_1 + \mathbf{M}^{-1}\mathbf{BQC}_0\|_\infty < \gamma$. Assuming that the necessary condition 1 holds, we therefore need to show that there exists $\mathbf{Q} \in \mathcal{RH}_\infty$ such that $\|\mathbf{M}^{-1}\mathbf{R}_1 + \mathbf{M}^{-1}\mathbf{BQC}_0\|_\infty < \gamma \Leftrightarrow$ there exists $\mathbf{W}$ satisfying (2.11). But, since $\mathbf{C}_0 \in \mathcal{GH}_\infty$, this is just a unilateral model matching problem. By Theorem 2.4 we know that $\mathbf{Q}$ exists if and only if there exists $\mathbf{Y} \in \mathcal{GH}_\infty$ with $\mathbf{Y}_{11} \in \mathcal{GH}_\infty$ such that

$$\mathbf{Y}^\sim J\mathbf{Y} = \mathbf{P}_1^\sim J\mathbf{P}_1, \qquad \mathbf{P}_1 = \begin{bmatrix} \mathbf{M}^{-1}\mathbf{B} & \mathbf{M}^{-1}\mathbf{R}_1 \\ 0 & I \end{bmatrix},$$

and that $\mathbf{Y}^{-1}$ "generates" all $\mathbf{QC}_0$'s. But such a $\mathbf{Y}$ exists if and only if there exists $\mathbf{W} \in \mathcal{GH}_\infty$ with $\mathbf{W}_{11} \in \mathcal{GH}_\infty$ satisfying

$$\mathbf{W}^{\sim} J \mathbf{W} = \mathbf{P}^{\sim} J \mathbf{P}, \qquad \mathbf{P} = \mathbf{P}_1 \begin{bmatrix} I & 0 \\ 0 & \mathbf{C}_0^{-1} \end{bmatrix},$$

and furthermore $\mathbf{W}^{-1}$ "generates" all $\mathbf{Q}$'s. It remains therefore to show that $\mathbf{P}^{\sim} J \mathbf{P} = \mathbf{G}^{\sim} J \mathbf{G}$, with $\mathbf{G}$ as in (2.11):

$$\begin{aligned} \mathbf{P} &= \begin{bmatrix} \mathbf{M}^{-1} & \mathbf{M}^{-1}\mathbf{R}_1\mathbf{C}_0^{-1} \\ 0 & \mathbf{C}_0^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{B} & 0 \\ 0 & I \end{bmatrix} \\ &= \hat{J} \begin{bmatrix} \mathbf{C}_0 & 0 \\ \mathbf{R}_1 & \mathbf{M} \end{bmatrix}^{-1} \hat{J}^* \begin{bmatrix} \boldsymbol{B} & 0 \\ 0 & I \end{bmatrix}. \end{aligned} \tag{2.14}$$

Now observe that $\hat{J}^* J \hat{J} = -\gamma^2 J^{-1}$, that $\hat{J}\hat{J}^* = I$ and that

$$\begin{bmatrix} \mathbf{C}_0 & 0 \\ \mathbf{R}_1 & \mathbf{M} \end{bmatrix} J \begin{bmatrix} \mathbf{C}_0 & 0 \\ \mathbf{R}_1 & \mathbf{M} \end{bmatrix}^{\sim} = \mathbf{HJH}^{\sim} = \mathbf{VJV}^{\sim}.$$

It is then easy to check that $\mathbf{G}^{\sim}\mathbf{JG} = \mathbf{P}^{\sim}\mathbf{JP}$. □

Remark 2.7. Suppose $\mathbf{V}$ as in part one of the Theorem exists and that $\mathbf{G}$ is as given in (2.11). Since $\mathbf{G}^{\sim} J \mathbf{G} = \mathbf{P}^{\sim} J \mathbf{P}$ with $\mathbf{P}$ as in (2.14), it follows that if $\mathbf{W}$ satisfies (2.11) then the condition $\mathbf{W}_{11}^* \mathbf{W}_{11} - \gamma^2 \mathbf{W}_{21}^* \mathbf{W}_{21} \geqq 0$ of Lemma 2.2 part two will be satisfied. Hence if *any* $\mathbf{W} \in \mathcal{GH}_\infty$ satisfying (2.11) has the property $\mathbf{W}_{11} \in \mathcal{GH}_\infty$, then *all* do.

3. *J*-spectral factorization theory. In the last section we solved the model matching problem in terms of J-spectral factorization. For the most part, the arguments made no reference to state space ideas. It is this connection that we now investigate. Specifically, we will relate the existence of J-spectral factors to the existence of solutions to indefinite algebraic Riccati equations. The main tool for this work is the state space factorization theory of Bart, Gohberg, and Kaashoek [6]. We begin with a little notation.

DEFINITION 3.1. A matrix $H \in \mathbb{C}^{2n \times 2n}$ is a Hamiltonian matrix if $\hat{J}H = H^*\hat{J}^*$, $\hat{J} = \left[\begin{smallmatrix} 0 & -I_n \\ I_n & 0 \end{smallmatrix}\right]$. If $H \in \mathbb{C}^{2n \times 2n}$ is a Hamiltonian matrix, we say $H \in \operatorname{dom}(\mathrm{Ric})$ if there exists $Q \in \mathbb{C}^{n \times n}$ and $\Lambda \in \mathbb{C}^{n \times n}$ such that

$$H \begin{bmatrix} I_n \\ Q \end{bmatrix} = \begin{bmatrix} I_n \\ Q \end{bmatrix} \Lambda$$

with Λ asymptotically stable (i.e., $\operatorname{In}(\Lambda) = (0, n, 0)$). If $H \in \operatorname{dom}(\mathrm{Ric})$, then $Q = \mathrm{Ric}(H)$ is Hermitian and satisfies the algebraic Riccati equation

$$QH_{11} + H_{11}^* Q + QH_{12}Q - H_{21} = 0$$

with

$$H_{11} + H_{12}Q = \Lambda \text{ asymptotically stable.}$$

We now prove the equivalence between J-spectral factorization and the solution of indefinite Riccati equations. A related result is in [5].

THEOREM 3.2. *Suppose* $\mathbf{G} \in \mathcal{RH}_\infty^{(p+q)\times(m+l)}$ *is given by the realization* $\mathbf{G}(s) = D + C(sI - A)^{-1}B$, *with* $A \in \mathbb{C}^{n \times n}$ *asymptotically stable* (*i.e.*, $\operatorname{In}(A) = (0, n, 0)$). *Then there exists a* $\mathbf{W} \in \mathcal{GH}_\infty$ *such that*

$$\mathbf{G}^{\sim} J_{pq}(\gamma) \mathbf{G} = \mathbf{W}^{\sim} \mathbf{J}_{ml}(\gamma) \mathbf{W} \tag{3.1}$$

if and only if:

1. *There exists a nonsingular matrix* $W_\infty \in \mathbb{C}^{(m+l)\times(m+l)}$ *such that*

$$D^* J_{pq}(\gamma) D = W_\infty^* J_{ml}(\gamma) W_\infty \tag{3.2}$$

and

2. $H \in \operatorname{dom}(\mathrm{Ric})$, *where*

$$H = \begin{bmatrix} A & 0 \\ -C^*JC & -A^* \end{bmatrix} - \begin{bmatrix} B \\ -C^*JD \end{bmatrix} (D^*JD)^{-1} [D^*JC \quad B^*]. \tag{3.3}$$

(*Here*, $J = J_{pq}(\gamma)$).

In this case $\mathbf{W} \in \mathscr{GH}_\infty$ *satisfies* (3.1) *if and only if, for some solution* W_∞ *of* (3.2), $\mathbf{W}$ *is given by*

$$\mathbf{W}(s) = W_\infty + L(sI - A)^{-1} B \tag{3.4a}$$

where

$$L = J_{ml}^{-1}(\gamma) W_\infty^{-*} (D^* J_{pq}(\gamma) C + B^* Q) \tag{3.4b}$$

$$Q = \mathrm{Ric}(H). \tag{3.4c}$$

Proof. Suppose 1 and 2 hold. Then $Q = \mathrm{Ric}(H)$ implies that $A - B(D^*JD)^{-1}[D^*JC + B^*Q] = A - BW_\infty^{-1}L$ is asymptotically stable. It follows, with $\mathbf{W}$ defined by (3.4), that $\mathbf{W} \in \mathscr{GH}_\infty$. Now note that the Riccati equation for Q can be written as

$$QA + A^*Q + C^*JC - L^*JL = 0 \tag{3.5}$$

with L as in (3.4b). Hence

$$\begin{aligned} \mathbf{W}^\sim J \mathbf{W} &= [W_\infty^* + B^*(-sI - A^*)^{-1} L^*] J [W_\infty + L(sI - A)^{-1} B] \\ &= D^*JD + [D^*JC + B^*Q](sI - A)^{-1}B + B^*(-sI - A^*)^{-1}[C^*JD + QB] \\ &\quad - B^*(-sI - A^*)^{-1}[Q(sI - A) + (-sI - A^*)Q - C^*JC](sI - A)^{-1}B \\ &= [D^* + B^*(-sI - A^*)^{-1}C^*] J [D + C(sI - A)^{-1}B] \\ &= \mathbf{G}^\sim J \mathbf{G}. \end{aligned}$$

That (3.4) gives all $\mathbf{W}$ follows from Lemma 2.2.

Now suppose there exists $\mathbf{W} \in \mathscr{GH}_\infty$ such that $\mathbf{G}^\sim J \mathbf{G} = \mathbf{W}^\sim J \mathbf{W}$. It follows by evaluating (3.1) at $s = \infty$ that (3.2) has a solution $W_\infty = \mathbf{W}(\infty)$. Let $\mathbf{M} = \mathbf{G}^\sim J \mathbf{G}$, $\mathbf{M}_+ = \mathbf{W}^\sim J$ and $\mathbf{M}_- = \mathbf{W}$. We then have $\mathbf{M} = \mathbf{M}_+ \mathbf{M}_-$, $\mathbf{M}_- \in \mathscr{GH}_\infty$, $\mathbf{M}_+^\sim \in \mathscr{GH}_\infty$, which is a canonical Wiener–Hopf factorization of $\mathbf{M}$. To establish that $H \in \operatorname{dom}(\mathrm{Ric})$, we use the factorization theorem of Bart, Gohberg, and Kaashoek [6] (see also [10, Chap. 7]). The relevant result is the following theorem.

THEOREM (BGK). *Suppose* $\mathbf{M} = \hat{D} + \hat{C}(sI - \hat{A})^{-1}\hat{B}$ *with* $(\hat{A}, \hat{B}, \hat{C})$ *minimal,* $\hat{A} \in \mathbb{C}^{n\times n}$. *Then* $\mathbf{M}$ *has a canonical Wiener–Hopf factorization if and only if* $\hat{D}$ *is invertible,* $\hat{A}$ *and* $\hat{A}^\times = \hat{A} - \hat{B}\hat{D}^{-1}\hat{C}$ *have no imaginary axis eigenvalues and* $X_+(\hat{A})$ *and* $X_-(\hat{A}^\times)$ *are complementary* (*i.e.*, $X_+(\hat{A}) \cap X_-(\hat{A}^\times) = \{0\}$ *and* $X_+(\hat{A}) \cup X_-(\hat{A}^\times) = \mathbb{C}^n$), *where* $X_+(\hat{A})$ (*respectively*, $X_-(\hat{A})$) *is the subspace of* $\mathbb{C}^n$ *spanned by the generalized eigenvectors of* $\hat{A}$ *corresponding to eigenvalues* λ *of* $\hat{A}$ *such that* $\mathrm{Re}(\lambda) > 0$ (*respectively*, $\mathrm{Re}(\lambda) < 0$).

The problem in applying this theorem in our case is that the realization of $\mathbf{M} = \mathbf{G}^\sim J \mathbf{G}$ is not required to be minimal under our assumptions. The assumption that A is asymptotically stable in the realization (3.1) allows us to avoid the minimality

condition by applying the BGK theorem to a minimal realization of $\mathbf{M} = \mathbf{G}^{\sim} J \mathbf{G}$ and then showing that the dilation to the original realization does not destroy the complementarity of the subspaces. We are going to do this in two steps: First we assume that (A, B) is controllable in the realization of G.

Temporary assumption. (A, B) controllable.

Since A is asymptotically stable, there exists $P = P^*$ (unique) such that

$$PA + A^*P + C^*JC = 0.$$

It follows that $\mathbf{G}^{\sim} J \mathbf{G}$ is given by

$$\mathbf{G}^{\sim} J \mathbf{G} \overset{s}{=} \left[\begin{array}{cc|c} A & 0 & B \\ 0 & -A^* & -K^* \\ \hline K & B^* & D^*JD \end{array}\right], \quad K = D^*JC + B^*P. \tag{3.6}$$

Since (A, B) is controllable, the unobservable (respectively, uncontrollable) modes of the realization (3.6) are the unobservable modes of (K, A) (respectively, uncontrollable modes of $(-A^*, -K^*)$).

Therefore, without loss of generality suppose A, B, C are such that

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad K = [K_1 \quad 0] \quad (K_1, A_{11}) \text{ observable.} \tag{3.7}$$

A minimal realization of $\mathbf{G}^{\sim} J \mathbf{G}$ is given by

$$\mathbf{G}^{\sim} J \mathbf{G} \overset{s}{=} \left[\begin{array}{cc|c} A_{11} & 0 & B_1 \\ 0 & -A_{11}^* & -K_1^* \\ \hline K_1 & B_1^* & D^*JD \end{array}\right] = \left[\begin{array}{c|c} \hat{A} & \hat{B} \\ \hline \hat{C} & \hat{D} \end{array}\right]. \tag{3.8}$$

By the BGK theorem, since $\mathbf{G}^{\sim} J \mathbf{G}$ has a canonical factorization, the Hamiltonian matrix $\hat{A}^{\times} = \hat{A} - \hat{B}\hat{D}^{-1}\hat{C}$ has no imaginary axis eigenvalues. Hence there exists nonsingular matrix $\hat{X}$ such that

$$\hat{A}^{\times}\hat{X} = \hat{X}T, \quad T = \begin{bmatrix} T_1 & T_2 \\ 0 & T_3 \end{bmatrix} \quad \operatorname{Re}\{\lambda_i(T_1)\} < 0, \operatorname{Re}\{\lambda_i(T_3)\} > 0 \; i = 1, \cdots, n. \tag{3.9}$$

Partition $\hat{X}$ conformably with T. We see from (3.8) and (3.9) that

$$X_+(\hat{A}) = \operatorname{Im}\begin{bmatrix} 0 \\ I \end{bmatrix} \quad \text{and} \quad X_-(\hat{A}^{\times}) = \operatorname{Im}\begin{bmatrix} \hat{X}_{11} \\ \hat{X}_{21} \end{bmatrix}.$$

By the BGK theorem $X_+(\hat{A})$ and $X_-(\hat{A}^{\times})$ are complementary, i.e.,

$$\begin{bmatrix} \hat{X}_{11} & 0 \\ \hat{X}_{21} & I \end{bmatrix} \quad \text{nonsingular.} \tag{3.10}$$

Hence $\hat{Q} = \hat{X}_{21}\hat{X}_{11}^{-1} = \operatorname{Ric}(\hat{A}^{\times})$.

Now return to the realization (3.6) with (A, B, K) as in (3.7). Consider

$$\begin{aligned} \tilde{H} &= \begin{bmatrix} A & 0 \\ 0 & -A^* \end{bmatrix} - \begin{bmatrix} B \\ -K^* \end{bmatrix} (D^*JD)^{-1} [K \quad B^*] \\ &= \begin{bmatrix} \hat{A}_{11}^{\times} & 0 & \hat{A}_{12}^{\times} & \tilde{H}_{14} \\ \tilde{H}_{21} & A_{22} & \tilde{H}_{23} & \tilde{H}_{24} \\ \hat{A}_{21}^{\times} & 0 & \hat{A}_{22}^{\times} & \tilde{H}_{34} \\ 0 & 0 & 0 & -A_{22}^* \end{bmatrix}. \end{aligned}$$

Observe that

$$\tilde{H}\begin{bmatrix}\hat{X}_{11} & 0\\ 0 & I\\ \hat{X}_{21} & 0\\ 0 & 0\end{bmatrix}=\begin{bmatrix}\hat{X}_{11} & 0\\ 0 & I\\ \hat{X}_{21} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}T_1 & 0\\ \tilde{T}_{21} & A_{22}\end{bmatrix}$$

and furthermore, with H as in (3.3) we have that $\left[\begin{smallmatrix} I & 0\\ -P & I\end{smallmatrix}\right]H\left[\begin{smallmatrix} I & 0\\ P & I\end{smallmatrix}\right]=\tilde{H}$. It follows that $H\in\operatorname{dom}(\mathrm{Ric})$ and $Q=\operatorname{Ric}(H)=\left[\begin{smallmatrix}\hat{Q} & 0\\ 0 & 0\end{smallmatrix}\right]+P$.

Removal of the controllability assumption. Suppose (A, B, C) is in controllable canonical form:

$$(3.11)\quad A=\begin{bmatrix}A_{11} & A_{12}\\ 0 & A_{22}\end{bmatrix},\qquad B=\begin{bmatrix}B_1\\ 0\end{bmatrix}\quad (A_{11}, B_1)\text{ controllable},\qquad C=[C_1\quad C_2]$$

and define $\tilde{H}$ by

$$(3.12)\qquad \tilde{H}=\begin{bmatrix}A_{11} & 0\\ -C_1^*JC_1 & -A_{11}^*\end{bmatrix}-\begin{bmatrix}B_1\\ -C_1^*JD\end{bmatrix}(D^*JD)^{-1}[D^*JC_1\ B_1^*].$$

Applying the above result (i.e., with the controllability assumption), we have $\tilde{H}\in\operatorname{dom}(\mathrm{Ric})$ and so there exists $\tilde{Q}$ such that

$$\tilde{H}\begin{bmatrix}I\\ \tilde{Q}\end{bmatrix}=\begin{bmatrix}I\\ \tilde{Q}\end{bmatrix}\tilde{\Lambda}\quad\text{with }\tilde{\Lambda}\text{ asymptotically stable (i.e., }\operatorname{In}(\tilde{\Lambda})=(0, n, 0)).$$

Now consider H defined by (3.3). Since (A, B, C) is in controllable canonical form, H is as follows:

$$H=\begin{bmatrix}\tilde{H}_{11} & H_{12} & \tilde{H}_{12} & 0\\ 0 & A_{22} & 0 & 0\\ \tilde{H}_{21} & H_{32} & -\tilde{H}_{11}^* & 0\\ H_{32}^* & H_{42} & -H_{12}^* & -A_{22}^*\end{bmatrix}.$$

Since $\tilde{\Lambda}=\tilde{H}_{11}+\tilde{H}_{12}\tilde{Q}$ and A_{22} are asymptotically stable, there exist Q_{12} and Q_{22} such that

$$\begin{aligned}&Q_{12}A_{22}+\tilde{\Lambda}^*Q_{12}=H_{32}-\tilde{Q}H_{12}\\ &Q_{22}A_{22}+A_{22}^*Q_{22}=H_{42}-H_{12}^*Q_{12}-Q_{12}^*(H_{12}+\tilde{H}_{12}Q_{12}),\end{aligned}$$

it follows that

$$H\begin{bmatrix}I & 0\\ 0 & I\\ \tilde{Q} & Q_{12}\\ Q_{12}^* & Q_{22}\end{bmatrix}=\begin{bmatrix}I & 0\\ 0 & I\\ \tilde{Q} & Q_{12}\\ Q_{12}^* & Q_{22}\end{bmatrix}\begin{bmatrix}\tilde{\Lambda} & H_{12}+\tilde{H}_{12}Q_{12}\\ 0 & A_{22}\end{bmatrix}$$

and we see that $H\in\operatorname{dom}(\mathrm{Ric})$.

4. State-space solution of the model matching problem. We are now ready to apply the J-spectral factorization results to the model matching problem associated via (1.6) with the standard $\mathscr{H}_\infty$ generalized regulator problem [8], [10], [23].

4.1. State-space preliminaries. Throughout the remainder of the paper we will assume that $\mathbf{P}(s)$ has state-space realization given by

$$\mathbf{P} \overset{s}{=} \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right] \tag{4.1}$$

where we assume:

A1. (A, B_2) is stabilizable and (C_2, A) is detectable.

A2. $D_{12}^* D_{12} = I$ and $D_{21} D_{21}^* = I$. We will also denote the unitary completions of D_{12} and D_{12} as $D_\perp$ and $\tilde{D}_\perp$.

As has already been noted [24], [13], the assumption implicit in (4.1) that $D_{11} = 0$, $D_{22} = 0$ can be made without loss of generality—by using a loop shifting argument which in the present context amounts to solving the factorization at ∞ problem first (see (3.2)) and introducing a (γ-dependent) change of variables. It is of course also possible to directly tackle the factorizations without assuming any special structure for D, but this considerably increases the length of the calculations.

By A1, there exist state feedback and output injection matrices F and H such that $A - B_2F$ and $A - HC_2$ are asymptotically stable. A doubly coprime factorization of $\mathbf{P}_{22}$, i.e.,

$$\mathbf{P}_{22} = \mathbf{N}_r \mathbf{D}_r^{-1} = \mathbf{D}_l^{-1} \mathbf{N}_l$$

with

$$\begin{bmatrix} \mathbf{V}_r & \mathbf{U}_r \\ -\mathbf{N}_l & \mathbf{D}_l \end{bmatrix} \begin{bmatrix} \mathbf{D}_r & -\mathbf{U}_l \\ \mathbf{N}_r & \mathbf{V}_l \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

is given by

$$\begin{bmatrix} \mathbf{D}_r & -\mathbf{U}_l \\ \mathbf{N}_r & \mathbf{V}_l \end{bmatrix} \overset{s}{=} \left[\begin{array}{c|cc} A - B_2F & B_2 & H \\ \hline -F & I & 0 \\ C_2 & 0 & I \end{array}\right]. \tag{4.2}$$

We then get the $\mathbf{T}_{ij}$'s of the associated model matching problem as [8], [10], [23]

$$\begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & 0 \end{bmatrix} \overset{s}{=} \left[\begin{array}{cc|cc} A - B_2F & B_2F & B_1 & B_2 \\ 0 & A - HC_2 & B_1 - HD_{21} & 0 \\ \hline C_1 - D_{12}F & D_{12}F & 0 & D_{12} \\ 0 & C_2 & D_{21} & 0 \end{array}\right]. \tag{4.3}$$

LEMMA 4.1. $\mathbf{T}_{21}$ *(respectively,* $\mathbf{T}_{12}$*) has a right (respectively, left) inverse in* $\mathcal{RL}_\infty$ *if and only if* $\left[\begin{smallmatrix} A-\lambda I & B_1 \\ C_2 & D_{21} \end{smallmatrix}\right]$ *has full row rank (respectively,* $\left[\begin{smallmatrix} A-\lambda I & B_2 \\ C_1 & D_{12} \end{smallmatrix}\right]$ *has full column rank) for all* $\lambda + \bar{\lambda} = 0$.

Proof. $\mathbf{T}_{21}$ right invertible in $\mathcal{RL}_\infty \Leftrightarrow \mathbf{T}_{21}(\lambda)$ full row rank for all $\lambda + \bar{\lambda} = 0$. Since $A - HC_2$ is asymptotically stable, $(A - HC_2 - \lambda I)$ is nonsingular for any $\lambda + \bar{\lambda} = 0$. Hence for $\lambda + \bar{\lambda} = 0$,

$$\begin{aligned} & u^* \mathbf{T}_{21}(\lambda) = 0, \qquad u \neq 0 \\ \Leftrightarrow & \begin{bmatrix} x^* & u^* \end{bmatrix} \begin{bmatrix} A - HC_2 - \lambda I & B_1 - HD_{21} \\ C_2 & D_{21} \end{bmatrix} = 0, \qquad x \neq 0, \qquad u \neq 0 \\ \Leftrightarrow & \begin{bmatrix} x^* & u^* \end{bmatrix} \begin{bmatrix} I & -H \\ 0 & I \end{bmatrix} \begin{bmatrix} A - \lambda I & B_1 \\ C_2 & D_{21} \end{bmatrix} = 0. \end{aligned}$$

□

4.2. A unilateral model matching problem. We now derive necessary and sufficient conditions, in terms of a nonnegative definiteness condition on the solution of an indefinite Riccati equation, for the existence of $\mathbf{Q} \in \mathcal{RH}_\infty$ such that $\|\mathbf{T}_{11} + \mathbf{QT}_{21}\|_\infty < \gamma$. We do this via Theorems 2.4 and 3.2. Consider $\mathbf{H}$ defined by

$$\mathbf{H} = \begin{bmatrix} \mathbf{T}_{21} & 0 \\ \mathbf{T}_{11} & I_l \end{bmatrix}. \tag{4.4}$$

By Theorem 2.4, applied to the matrix $\mathbf{G}(s) = \mathbf{H}(\bar{s})^*$, we need to solve the following factorization problem.

FACTORIZATION PROBLEM P1. *With* $\mathbf{H} \in \mathcal{RH}_\infty^{(m+l)\times(p+l)}$ *defined by* (4.4), *find* $\mathbf{V} \in \mathcal{GH}_\infty^{m+l}$ *with* $\mathbf{V}_{11} \in \mathcal{GH}_\infty^m$ *such that*

$$\mathbf{H}J_{pl}(\gamma)\mathbf{H}^\sim = \mathbf{V}J_{ml}(\gamma)\mathbf{V}^\sim. \tag{4.5}$$

THEOREM 4.2. *Let* $\mathbf{H}$ *be as in* (4.4). *Then Problem* P1 *has a solution if and only if* $H_Y \in \operatorname{dom}(\mathrm{Ric})$ *and* $\mathrm{Ric}(H_Y) \geqq 0$, *where*

$$H_Y = \begin{bmatrix} A^* & 0 \\ -B_1B_1^* & -A \end{bmatrix} - \begin{bmatrix} C_2^* & C_1^* \\ -B_1D_{21}^* & 0 \end{bmatrix} J^{-1} \begin{bmatrix} D_{21}B_1^* & C_2 \\ 0 & C_1 \end{bmatrix}. \tag{4.6}$$

($J = J_{pl}(\gamma)$). *In this case, a solution* $\mathbf{V}$ *to Problem* P1 *is given by*

$$\mathbf{V} = \begin{bmatrix} \mathbf{D}_l & 0 \\ -\mathbf{P}_{12}\mathbf{U}_l\mathbf{D}_l & I \end{bmatrix} \mathbf{V}_1 \tag{4.7}$$

where

$$\mathbf{V}_1 \overset{s}{=} \left[\begin{array}{c|cc} A & M_1 & M_2 \\ \hline C_2 & I_m & 0 \\ C_1 & 0 & I_l \end{array}\right] \tag{4.8a}$$

and

$$M = [M_1 \quad M_2] = [\,Y_\infty C_2^* + B_1D_{21}^* - \gamma^{-2}Y_\infty C_1^*] \tag{4.8b}$$

with

$$Y_\infty = \mathrm{Ric}(H_Y). \tag{4.8c}$$

Proof. Write $\mathbf{H}$ as

$$\mathbf{H} = \mathbf{H}_1\mathbf{H}_2 \tag{4.9}$$

where

$$\mathbf{H}_1 \overset{s}{=} \left[\begin{array}{c|cc} A - B_2F & H & 0 \\ \hline 0 & I & 0 \\ C_1 - D_{12}F & 0 & I \end{array}\right] \tag{4.10a}$$

$$\mathbf{H}_2 \overset{s}{=} \left[\begin{array}{c|cc} A - HC_2 & B_1 - HD_{21} & 0 \\ \hline C_2 & D_{21} & 0 \\ C_1 & 0 & I \end{array}\right]. \tag{4.10b}$$

Since $\mathbf{H}_1 \in \mathcal{GH}_\infty$ and has the particular form $\mathbf{H}_1 = \left[\begin{smallmatrix} I & 0 \\ \mathbf{X} & I \end{smallmatrix}\right]$, we see that $\mathbf{V}$ solves Problem P1 if and only if there exists $\mathbf{V}_2 \in \mathcal{GH}_\infty$ with $(\mathbf{V}_2)_{11} \in \mathcal{GH}_\infty$ such that $\mathbf{H}_2J\mathbf{H}_2^\sim = \mathbf{V}_2J\mathbf{V}_2^\sim$;

$\mathbf{V}$ and $\mathbf{V}_2$ are related via $\mathbf{V}=\mathbf{H}_1\mathbf{V}_2$. Applying Theorem 3.2, we see that $H_Y \in \operatorname{dom}(\mathrm{Ric})$ is necessary and sufficient for the existence of $\mathbf{V}_2 \in \mathscr{GH}_\infty$, and that $\mathbf{V}_2$ is given by

$$\mathbf{V}_2 \overset{s}{=} \left[\begin{array}{c|cc} A-HC_2 & M_1-H & M_2 \\ \hline C_2 & I_m & 0 \\ C_1 & 0 & I_l \end{array}\right] \tag{4.11}$$

with M as in (4.8b).

We now claim $(\mathbf{V}_2)_{11} \in \mathscr{GH}_\infty \Leftrightarrow Y_\infty = \mathrm{Ric}(H_Y) \geqq 0$. Since $(A-HC_2)$ is asymptotically stable, it follows that $(\mathbf{V}_2)_{11} \in \mathscr{GH}_\infty \Leftrightarrow A-M_1C_2$ is asymptotically stable. We therefore need to show that $A-M_1C_2$ is asymptotically stable $\Leftrightarrow Y_\infty \geqq 0$. To see this, write the Riccati equation for Y_∞ as

$$AY_\infty + Y_\infty A^* + B_1B_1^* - MJM^* = 0. \tag{4.12}$$

Since $M_1 = Y_\infty C_2^* + B_1D_{21}^*$ we see that

$$M_1M_1^* = Y_\infty C_2^*M_1^* + M_1C_2Y_\infty - Y_\infty C_2^*C_2Y_2 + B_1D_{21}^*D_{21}B_1^*. \tag{4.13}$$

Substituting into (4.12) we obtain

$$(A-M_1C_2)Y_\infty + Y_\infty(A-M_1C_2)^* + [\,Y_\infty C_2^* \quad \gamma M_2 \quad B_1\tilde{D}_\perp^*] \begin{bmatrix} C_2Y_\infty \\ \gamma M_2^* \\ \tilde{D}_\perp B_1^* \end{bmatrix} = 0. \tag{4.14}$$

Since $(A-M_1C_2-M_2C_1)$ is asymptotically stable, $(A-M_1C_2, M_2)$ is stabilizable. Hence [26, Lemma 12.2], $Y_\infty \geqq 0 \Leftrightarrow (A-M_1C_2)$ is asymptotically stable.

It remains to verify the formula (4.7) for $\mathbf{V}=\mathbf{H}_1\mathbf{V}_2$. This is easily done via a state space calculation. □

Remark 4.3. The decomposition of $\mathbf{V}$ in (4.7) is analogous to the decomposition of $\mathbf{H}$ as

$$\mathbf{H} = \begin{bmatrix} \mathbf{D}_l & 0 \\ -\mathbf{P}_{12}\mathbf{U}_l\mathbf{D}_l & I \end{bmatrix} \begin{bmatrix} \mathbf{P}_{21} & 0 \\ \mathbf{P}_{11} & I \end{bmatrix} \tag{4.15}$$

(see (1.6)). It follows that $\mathbf{V}_1$ is a solution to the J-factorization observed in [12], namely

$$\mathbf{V}_1 J \mathbf{V}_1^{\sim} = \begin{bmatrix} \mathbf{P}_{21} & 0 \\ \mathbf{P}_{11} & I \end{bmatrix} J \begin{bmatrix} \mathbf{P}_{21} & 0 \\ \mathbf{P}_{11} & I \end{bmatrix}^{\sim}. \tag{4.16}$$

Remark 4.4. A necessary condition for $H_Y \in \operatorname{dom}(\mathrm{Ric})$ is that H_Y have no imaginary axis eigenvalues. It is not difficult to show that a necessary condition for this is that $\left[\begin{smallmatrix} A-\lambda I & B_1 \\ C_2 & D_{21} \end{smallmatrix}\right]$ be full row rank for all $\lambda+\bar{\lambda}=0$, since

$$[x_1^* \quad x_2^*] \begin{bmatrix} A-\lambda I & B_1 \\ C_2 & D_{21} \end{bmatrix} = 0 \Rightarrow x_1^*(A-B_1D_{21}^*C_2) = 0 \quad \text{and} \quad x_1^*B_1(I-D_{21}^*D_{21}) = 0$$

$$\Rightarrow [0 \quad x_1^*]H_Y = \lambda[0 \quad x_1^*].$$

An alternative view of this necessary condition is obtained by considering the J-spectral factorization directly, since a necessary condition for the factorization (4.5) to exist (with $\mathbf{V} \in \mathscr{GH}_\infty$) is that $\mathbf{H}$ (equivalently $\mathbf{T}_{21}$) be right invertible in $\mathscr{RL}_\infty$. This is equivalent to $\left[\begin{smallmatrix} A-\lambda I & B_1 \\ C_2 & D_{21} \end{smallmatrix}\right]$ full row rank for all $\lambda+\bar{\lambda}$ by Lemma 4.1.

Remark 4.5. The problem of finding $\mathbf{Q} \in \mathcal{RH}_\infty$ such that $\|\mathbf{T}_{11}+\mathbf{T}_{12}\mathbf{Q}\|_\infty < \gamma$ can be tackled in an entirely analogous way, applying Theorem 3.2 to the matrix

$$\mathbf{E} = \begin{bmatrix} \mathbf{T}_{12} & \mathbf{T}_{11} \\ 0 & I \end{bmatrix}. \tag{4.17}$$

The relevant conditions are:

1. $H_X \in \operatorname{dom}(\mathrm{Ric})$, where

$$H_X = \begin{bmatrix} A & 0 \\ -C_1^* C_1 & -A^* \end{bmatrix} - \begin{bmatrix} B_2 & B_1 \\ -C_1^* D_{12} & 0 \end{bmatrix} J^{-1} \begin{bmatrix} D_{12}^* C_1 & B_2^* \\ 0 & B_1^* \end{bmatrix}. \tag{4.18}$$

2. $X_\infty = \operatorname{Ric}(H_X) \geqq 0$.

The factorization dual to (4.16), i.e.,

$$\begin{bmatrix} \mathbf{P}_{12} & \mathbf{P}_{11} \\ 0 & I \end{bmatrix}^{\sim} J \begin{bmatrix} \mathbf{P}_{12} & \mathbf{P}_{11} \\ 0 & I \end{bmatrix} = \mathbf{X}^{\sim} J \mathbf{X}, \qquad \mathbf{X} \in \mathcal{GH}_\infty, \qquad \mathbf{X}_{11} \in \mathcal{GH}_\infty \tag{4.19}$$

is the factorization associated with the $\mathcal{H}_\infty$ state feedback problem in [22], where $\mathbf{P}$ is assumed stable.

4.3. A bilateral model matching problem. We derive necessary and sufficient conditions, in terms of nonnegative definiteness conditions on the solutions of two indefinite Riccati equations, for the existence of $\mathbf{Q} \in \mathcal{RH}_\infty$ such that $\|\mathbf{T}_{11}+\mathbf{T}_{12}\mathbf{Q}\mathbf{T}_{21}\|_\infty < \gamma$. The first Riccati equation is associated with the factorization Problem P1 in § 4.2 (see (4.6)), which we will, in this section, assume has a solution. The second Riccati equation is associated with the factorization of the matrix

$$\mathbf{G} = \hat{J}\mathbf{V}^{-1}\hat{J}^* \begin{bmatrix} \mathbf{T}_{12} & 0 \\ 0 & I_m \end{bmatrix}. \tag{4.20}$$

By Theorem 2.6, we need to solve the following factorization problem.

FACTORIZATION PROBLEM P2. *With* $\mathbf{G}$ *defined by* (4.20), *find* $\mathbf{W} \in \mathcal{GH}_\infty^{q+m}$ *with* $\mathbf{W}_{11} \in \mathcal{GH}_\infty^q$ *such that*

$$\mathbf{G}^{\sim} J_{lm}(\gamma) \mathbf{G} = \mathbf{W}^{\sim} J_{qm}(\gamma) \mathbf{W}. \tag{4.21}$$

THEOREM 4.6. *Let* $\mathbf{G}$ *be as in* (4.20). *Then Problem* P2 *has a solution if and only if* $H_Z \in \operatorname{dom}(\mathrm{Ric})$ *and* $\operatorname{Ric}(H_Z) \geqq 0$, *where*

$$H_Z = \begin{bmatrix} A - M_2 C_1 & 0 \\ -C_1^* C_1 & -(A - M_2 C_1)^* \end{bmatrix} - \begin{bmatrix} B_2 - M_2 D_{12} & M_1 \\ -C_1^* D_{12} & 0 \end{bmatrix} J^{-1} \times \begin{bmatrix} D_{12}^* C_1 & (B_2 - M_2 D_{12})^* \\ 0 & M_1^* \end{bmatrix}. \tag{4.22}$$

($J = J_{lm}(\gamma)$). *In this case,* $\mathbf{W}$ *is given by*

$$\mathbf{W} = \mathbf{W}_1 \begin{bmatrix} \mathbf{D}_r & -\mathbf{U}_l \\ \mathbf{N}_r & \mathbf{V}_l \end{bmatrix} \tag{4.23a}$$

where

$$\mathbf{W}_1 \overset{s}{=} \left[\begin{array}{c|cc} A - M_1 C_2 - M_2 C_1 & B_2 - M_2 D_{12} & M_1 \\ \hline L_1 & I & 0 \\ L_2 & 0 & I \end{array}\right] \tag{4.23b}$$

and

$$L = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} = \begin{bmatrix} D_{12}^* C_1 + (B_2 - M_2 D_{12})^* Z_\infty \\ -(C_2 + \gamma^{-2} M_1^* Z_\infty) \end{bmatrix} \tag{4.24a}$$

with

$$Z_\infty = \mathrm{Ric}\,(H_Z). \tag{4.24b}$$

Proof. First, consider the formula for $\mathbf{G}$ in light of the fact that $\mathbf{V}$ is given by (4.7).

$$\begin{aligned}
\mathbf{G} &= \hat{J}\mathbf{V}^{-1}\hat{J}^* \begin{bmatrix} \mathbf{T}_{12} & 0 \\ 0 & I \end{bmatrix} \\
&= \hat{J}\mathbf{V}_1^{-1} \begin{bmatrix} \mathbf{D}_l^{-1} & 0 \\ \mathbf{P}_{12}\mathbf{U}_l & I \end{bmatrix} \hat{J}^* \begin{bmatrix} \mathbf{T}_{12} & 0 \\ 0 & I \end{bmatrix} \quad \text{by (4.7)} \\
&= \hat{J}\mathbf{V}_1^{-1}\hat{J}^* \begin{bmatrix} I & -\mathbf{P}_{12}\mathbf{U}_l \\ 0 & \mathbf{D}_l^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{P}_{12}\mathbf{D}_r & 0 \\ 0 & I \end{bmatrix}, \quad \text{since } \mathbf{T}_{12} = \mathbf{P}_{12}\mathbf{D}_r \\
&= \hat{J}\mathbf{V}_1^{-1}\hat{J}^* \begin{bmatrix} \mathbf{P}_{12} & 0 \\ -\mathbf{P}_{22} & I \end{bmatrix} \begin{bmatrix} \mathbf{D}_r & -\mathbf{U}_l \\ \mathbf{N}_r & \mathbf{V}_l \end{bmatrix}, \quad \text{using (1.4).}
\end{aligned}$$

Thus,

$$\mathbf{G} = \mathbf{G}_1 \begin{bmatrix} \mathbf{D}_r & -\mathbf{U}_l \\ \mathbf{N}_r & \mathbf{V}_l \end{bmatrix} \tag{4.25}$$

where

$$\mathbf{G}_1 = \hat{J}\mathbf{V}_1^{-1}\hat{J}^* \begin{bmatrix} \mathbf{P}_{12} & 0 \\ -\mathbf{P}_{22} & I \end{bmatrix} \tag{4.26a}$$

$$\overset{s}{=} \left[\begin{array}{cc|cc} A - M_1C_2 - M_2C_1 & & B_2 - M_2D_{12} & M_1 \\ \hline C_1 & & D_{12} & 0 \\ -C_2 & & 0 & I \end{array}\right]. \tag{4.26b}$$

Since $\begin{bmatrix} \mathbf{D}_r & -\mathbf{U}_l \\ \mathbf{N}_r & \mathbf{V}_l \end{bmatrix} \in \mathscr{GH}_\infty$ there exists $\mathbf{W} \in \mathscr{GH}_\infty$ such that $\mathbf{G}^\sim J\mathbf{G} = \mathbf{W}^\sim J\mathbf{W}$ if and only if $\mathbf{W}$ is given by (4.23a), where $\mathbf{W}_1 \in \mathscr{GH}_\infty$ satisfies

$$\mathbf{G}_1^\sim J\mathbf{G}_1 = \mathbf{W}_1 J\mathbf{W}_1. \tag{4.27}$$

Using the realization (4.26b) and Theorem 3.2, there exists $\mathbf{W}_1 \in \mathscr{GH}_\infty$ satisfying (4.27) if and only if $H_Z \in \mathrm{dom}\,(\mathrm{Ric})$, and in this case, $\mathbf{W}_1$ given by (4.23b) satisfies (4.27).

Let us now consider necessary and sufficient conditions for $\mathbf{W}_{11} \in \mathscr{GH}_\infty$.

Using (4.23), (4.2) and the state transformation $\begin{bmatrix} I & 0 \\ -I & I \end{bmatrix}$ the following realization for $\mathbf{W}$ is obtained:

$$\mathbf{W} \overset{s}{=} \left[\begin{array}{c|c} \hat{A} & \hat{B} \\ \hline \hat{C} & I \end{array}\right] \tag{4.28a}$$

where

$$\hat{A} = \begin{bmatrix} A - M_2C_1 - (B_2 - M_2D_{12})F & -(B_2 - M_2D_{12})F + M_1C_2 \\ M_2(C_1 - D_{12}F) & A - M_2D_{12}F - M_1C_2 \end{bmatrix} \tag{4.28b}$$

$$\hat{B} = \begin{bmatrix} B_2 - M_2D_{12} & M_1 \\ M_2D_{12} & H - M_1 \end{bmatrix} \tag{4.28c}$$

$$\hat{C} = \begin{bmatrix} L_1 - F & -F \\ L_2 + C_2 & C_2 \end{bmatrix}. \tag{4.28d}$$

The "A" matrix of $\mathbf{W}_{11}^{-1}$ is therefore

$$\tilde{A}=\begin{bmatrix} A-M_2C_1-(B_2-M_2D_{12})L_1 & M_1C_2 \\ M_2(C_1-D_{12}L_1) & A-M_1C_2 \end{bmatrix}. \tag{4.29}$$

Rewrite the Riccati equation for Z_∞ as:

$$\begin{aligned} &Z_\infty[A-M_2C_1-(B_2-M_2D_{12})L_1]+[A-M_2C_1-(B_2-M_2D_{12})L_1]^*Z_\infty \\ &\quad+Z_\infty[(B_2-M_2D_{12})(B_2-M_2D_{12})^*+\gamma^{-2}M_1M_1^*]Z_\infty+C_1^*(I-D_{12}D_{12}^*)C_1=0. \end{aligned} \tag{4.30}$$

Using (4.14), (4.30), and $M_2=-\gamma^{-2}Y_\infty C_1^*$, we therefore have

$$\begin{aligned} &\begin{bmatrix} Z_\infty & 0 \\ 0 & \gamma^2 I \end{bmatrix}\tilde{A}\begin{bmatrix} I & 0 \\ 0 & Y_\infty \end{bmatrix}+\begin{bmatrix} I & 0 \\ 0 & Y_\infty \end{bmatrix}\tilde{A}^*\begin{bmatrix} Z_\infty & 0 \\ 0 & \gamma^2 I \end{bmatrix} \\ &\quad=-\begin{bmatrix} Z_\infty & 0 \\ 0 & Y_\infty \end{bmatrix}\begin{bmatrix} B_2-M_2D_{12} & \gamma^{-1}M_1 \\ -C_1^*D_{12} & -\gamma C_2^* \end{bmatrix} \\ &\quad\times\begin{bmatrix} B_2-M_2D_{12} & \gamma^{-1}M_1 \\ -C_1^*D_{12} & -\gamma C_2^* \end{bmatrix}^*\begin{bmatrix} Z_\infty & 0 \\ 0 & Y_\infty \end{bmatrix} \\ &\quad-\begin{bmatrix} C_1^* \\ Y_\infty C_1^* \end{bmatrix}D_\perp D_\perp^*[C_1 \;\; C_1Y_\infty]-\begin{bmatrix} 0 \\ \gamma B_1 \end{bmatrix}\tilde{D}_\perp^*\tilde{D}_\perp[0 \;\; \gamma B_1^*]. \end{aligned} \tag{4.31}$$

Temporary assumption. Y_∞ nonsingular. With Y_∞ nonsingular, define

$$\tilde{Z}_\infty=\begin{bmatrix} Z_\infty & 0 \\ 0 & \gamma^2 Y_\infty^{-1} \end{bmatrix}. \tag{4.32}$$

Since $\hat{A}-\hat{B}\hat{C}$ is asymptotically stable, $([L_2+C_2 \;\; C_2], \tilde{A})$ is detectable. Observing that $L_2+C_2=-\gamma^{-2}M_1^*Z_\infty$ it follows from (4.31) and [26, Lemma 12.2] that $\tilde{A}$ is asymptotically stable $\Leftrightarrow \tilde{Z}_\infty \geqq 0$.

Removal of temporary assumption. Suppose, without loss of generality, the realization (A, B, C) is such that Y_∞ is of the form

$$Y_\infty=\begin{bmatrix} \hat{Y}_\infty & 0 \\ 0 & 0 \end{bmatrix}, \quad \hat{Y}_\infty \text{ nonsingular.}$$

It follows from (4.14) that $A-M_1C_2$ is upper triangular:

$$A-M_1C_2=\begin{bmatrix} X_{11} & X_{12} \\ 0 & X_{22} \end{bmatrix}, \quad X_{22} \text{ asymptotically stable.}$$

Furthermore, we see from (4.29), since $M_2=-\gamma^{-2}Y_\infty C_1^*$, that $\tilde{A}$ is also upper triangular:

$$\tilde{A}=\begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & X_{22} \end{bmatrix}.$$

Applying the Y_∞ nonsingular argument to the 1, 1 block gives $\tilde{A}_{11}$ asymptotically stable $\Leftrightarrow Z_\infty \geqq 0$, and hence $\tilde{A}$ is asymptotically stable $\Leftrightarrow Z_\infty \geqq 0$.

Remark 4.7. The structure (4.23a) of $\mathbf{W}$ is of great significance, as we now explain. Recall from Theorem 2.6 that all matrices $\mathbf{Q} \in \mathcal{RH}_\infty$ such that $\|\mathbf{T}_{11}+\mathbf{T}_{12}\mathbf{Q}\mathbf{T}_{21}\|_\infty \leqq \gamma$ are given by

$$\mathbf{Q}=\mathbf{Q}_1\mathbf{Q}_2^{-1}, \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix}=\mathbf{W}^{-1}\begin{bmatrix} \mathbf{U} \\ I \end{bmatrix} \quad \mathbf{U} \in \mathcal{RH}_\infty \text{ with } \|\mathbf{U}\|_\infty \leqq \gamma$$

where $\mathbf{W}$ solves Problem P2. Also recall, from (1.5), that all stabilizing controllers are

given by

$$\mathbf{K}=\mathbf{K}_1\mathbf{K}_2^{-1},\begin{bmatrix}\mathbf{K}_1\\ \mathbf{K}_2\end{bmatrix}=\begin{bmatrix}\mathbf{D}_r & -\mathbf{U}_l\\ \mathbf{N}_r & \mathbf{V}_l\end{bmatrix}\begin{bmatrix}\mathbf{Q}\\ I_m\end{bmatrix}\quad \mathbf{Q}\in\mathcal{RH}_\infty^{q\times m}.$$

It follows from (4.23a) that all stabilizing controllers $\mathbf{K}$ such that $\|\mathcal{F}(\mathbf{P},\mathbf{K})\|_\infty\leqq\gamma$ are given by

$$\mathbf{K}=\mathbf{K}_1\mathbf{K}_2^{-1},\begin{bmatrix}\mathbf{K}_1\\ \mathbf{K}_2\end{bmatrix}=\mathbf{W}_1^{-1}\begin{bmatrix}\mathbf{U}\\ I\end{bmatrix},\ \mathbf{U}\in\mathcal{RH}_\infty \text{ with } \|\mathbf{U}\|_\infty\leqq\gamma. \tag{4.33}$$

5. The controller generator. Theorem 4.6 gives necessary and sufficient conditions for internally stabilizing controllers $\mathbf{K}$ such that $\|\mathcal{F}(\mathbf{P},\mathbf{K})\|_\infty<\gamma$ to exist. Furthermore, (4.23b) and (4.33) provide a representation formula for all such controllers. The result we give in this section provides an alternative formula for controllers; there will be two changes. First, we will replace Z_∞ by an equivalent expression, since $Z_\infty=X_\infty(I-\gamma^{-2}Y_\infty X_\infty)^{-1}$, and second, we will transform the formula (4.33) into an equivalent feedback form more typical in the engineering literature.

THEOREM 5.1. *Suppose* $\mathbf{P}(s)$ *is given by the realization* (4.1), *that assumptions* A1 *and* A2 *hold and that*

A3.

$$\begin{bmatrix}A-\lambda I & B_2\\ C_1 & D_{12}\end{bmatrix}\quad\text{and}\quad\begin{bmatrix}A-\lambda I & B_1\\ C_2 & D_{21}\end{bmatrix}$$

are, respectively, full column and row rank for all $\lambda+\bar{\lambda}=0$. *Then there exists a rational matrix* $\mathbf{K}$ *such that* $\mathcal{F}(\mathbf{P},\mathbf{K})$ *is internally stable and* $\|\mathcal{F}(\mathbf{P},\mathbf{K})\|_\infty<\gamma$ *if and only if* $H_X\in\operatorname{dom}(\mathrm{Ric})$, $H_Y\in\operatorname{dom}\mathrm{Ric})$ *and*

$$X_\infty\geqq 0,\quad Y_\infty\geqq 0\quad\text{and}\quad\lambda_{\max}(X_\infty Y_\infty)<\gamma^2 \tag{5.1a}$$

where

$$X_\infty=\mathrm{Ric}(H_X),\qquad Y_\infty=\mathrm{Ric}(H_Y) \tag{5.1b}$$

with H_Y *and* H_X *as in* (4.6) *and* (4.18).

Furthermore, when the conditions (5.1) *hold, all controllers* $\mathbf{K}$ *such that* $\mathcal{F}(\mathbf{P},\mathbf{K})$ *is internally stable and* $\|\mathcal{F}(\mathbf{P},\mathbf{K})\|_\infty\leqq\gamma$ *are given by*

$$\mathbf{K}=\mathcal{F}(\mathbf{K}_a,\mathbf{U})\qquad\mathbf{U}\in\mathcal{RH}_\infty\quad\textit{with}\quad\|\mathbf{U}\|_\infty\leqq\gamma \tag{5.2}$$

where

$$\mathbf{K}_a\overset{s}{=}\left[\begin{array}{c|cc}A_k & B_{k1} & B_{k2}\\ \hline C_{k1} & 0 & I\\ C_{k2} & I & 0\end{array}\right] \tag{5.3a}$$

with

$$B_k=[\,Y_\infty C_2^*+B_1D_{21}^*\quad B_2+\gamma^{-2}Y_\infty C_1^*D_{12}] \tag{5.3b}$$

$$C_k=\begin{bmatrix}-(D_{12}^*C_1+B_2^*X_\infty)\\ -(C_2+\gamma^{-2}D_{21}B_1^*X_\infty)\end{bmatrix}(I-\gamma^{-2}Y_\infty X_\infty)^{-1} \tag{5.3c}$$

$$A_k=A-B_{k1}C_2+\gamma^{-2}Y_\infty C_1^*C_1+B_{k2}C_{k1}. \tag{5.3d}$$

Proof. We have already proved that γ-suboptimal controllers $\mathbf{K}$ exist $\Leftrightarrow\mathbf{Q}\in\mathcal{RH}_\infty$ exists such that $\|\mathbf{T}_{11}+\mathbf{T}_{12}\mathbf{Q}\mathbf{T}_{21}\|_\infty<\gamma\Leftrightarrow H_Y$ and $H_Z\in\operatorname{dom}(\mathrm{Ric})$ with $Y_\infty\geqq 0$ and $Z_\infty\geqq 0$ (provided $\mathbf{T}_{21}$ and $\mathbf{T}_{12}$ have right and left inverses in $\mathcal{RL}_\infty$, which is assured by Lemma 4.1 and A3).

We need to show, given $H_Y \in \text{dom}(\text{Ric})$ and $Y_\infty = \text{Ric}(H_Y) \geqq 0$, that $H_Z \in \text{dom}(\text{Ric})$ and $Z_\infty = \text{Ric}(H_Z) \geqq 0 \Leftrightarrow H_X \in \text{dom}(\text{Ric})$. $X_\infty = \text{Ric}(H_X) \geqq 0$ and $\lambda_{\max}(X_\infty Y_\infty) < \gamma^2$.

Observe that

$$\begin{bmatrix} I & \gamma^{-2} Y_\infty \\ 0 & I \end{bmatrix} H_Z \begin{bmatrix} I & -\gamma^{-2} Y_\infty \\ 0 & I \end{bmatrix} = H_X. \tag{5.4}$$

Suppose $H_X \in \text{dom}(\text{Ric})$, $X_\infty = \text{Ric}(H_X) \geqq 0$ and $\lambda_{\max}(X_\infty Y_\infty) < \gamma^2$. Then $(I - \gamma^{-2} Y_\infty X_\infty)$ is nonsingular, and from (5.4) we see that $H_Z \in \text{dom}(\text{Ric})$, with $Z_\infty = \text{Ric}(H_Z) = X_\infty (I - \gamma^{-2} Y_\infty X_\infty)^{-1}$. To see that $Z_\infty \geqq 0$, note that

$$Z_\infty (\gamma^{-2} Y_\infty X_\infty - I) + (\gamma^{-2} Y_\infty X_\infty - I)^* Z_\infty + (X_\infty + X_\infty) = 0.$$

It follows [11, Thm. 3.3, part 3] that $Z_\infty \geqq 0$, since $(\gamma^{-2} Y_\infty X_\infty - I)$ is asymptotically stable.

Conversely, suppose $H_Z \in \text{dom}(\text{Ric})$ and $Z_\infty = \text{Ric}(H_Z) \geqq 0$. Hence $(I + \gamma^{-2} Z_\infty Y_\infty)$ is nonsingular and from (5.4), $H_X \in \text{dom}(\text{Ric})$ with

$$X_\infty = \text{Ric}(H_X) = (I + \gamma^{-2} Z_\infty Y_\infty)^{-1} Z_\infty = Z_\infty (I + \gamma^{-2} Y_\infty Z_\infty)^{-1}.$$

Clearly $X_\infty \geqq 0$ and we see that $\lambda_{\max}(X_\infty Y_\infty) < \gamma^2$ since

$$\lambda_i(X_\infty Y_\infty) = \lambda_i\{(I + \gamma^{-2} Z_\infty Y_\infty)^{-1} Z_\infty Y_\infty\} = \gamma^2 \frac{\lambda_i(Z_\infty Y_\infty)}{\gamma^2 + \lambda_i(Z_\infty Y_\infty)}.$$

This concludes the proof of the necessary and sufficient conditions for the existence of $\mathbf{K}$.

By Remark 4.7, $\mathbf{K}$ is given by

$$\mathbf{K} = \mathbf{K}_1 \mathbf{K}_2^{-1}, \quad \begin{bmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \end{bmatrix} = \mathbf{W}_1^{-1} \begin{bmatrix} \mathbf{U} \\ I \end{bmatrix}, \qquad \mathbf{U} \in \mathcal{RH}_\infty \text{ with } \|\mathbf{U}\|_\infty < \gamma. \tag{5.5}$$

Defining $\mathbf{X} = \mathbf{W}_1^{-1}$, we can equivalently write

$$\mathbf{K} = \mathcal{F}(\mathbf{K}_a, \mathbf{U}), \qquad \mathbf{U} \in \mathcal{RH}_\infty \quad \text{with} \quad \|\mathbf{U}\|_\infty < \gamma$$

where

$$\mathbf{K}_a = \begin{bmatrix} \mathbf{X}_{12} & \mathbf{X}_{11} \\ I & 0 \end{bmatrix} \begin{bmatrix} \mathbf{X}_{22} & \mathbf{X}_{21} \\ 0 & I \end{bmatrix}^{-1}. \tag{5.6}$$

Rewrite L in (4.24a) as

$$L = \begin{bmatrix} D_{12}^* C_1 + B_2^* X_\infty \\ -(C_2 + \gamma^{-2} D_{21} B_1^* X_\infty) \end{bmatrix} (I - \gamma^{-2} Y_\infty X_\infty)^{-1}. \tag{5.7}$$

A straightforward state space calculation using (4.23) and (5.7) will reveal that a realization for $\mathbf{K}_a$ in (5.6) is indeed given by (5.3). □

We note here that this theorem agrees with others derived recently, such as [9], [12], [13], [21].

6. Conclusion. In this paper the J-spectral factorization approach to suboptimal $\mathcal{H}_\infty$ control problems of "Nehari"/"one-block"/"first kind" type has been extended to the general case.

The existence of solutions was shown to be equivalent to the existence of solutions to two coupled J-spectral factorization problems with the additional property that the $(1, 1)$ block of both J-spectral factors be outer. The second of these J-spectral factors was shown to generate all solutions to the $\mathcal{H}_\infty$ control problem.

The existence of the J-spectral factors was then shown to be equivalent to the existence of nonnegative definite, stabilizing solutions to two indefinite algebraic Riccati equations. This allowed an explicit state space formula for a generator of all solutions to the suboptimal $\mathcal{H}_\infty$ control problem to be given.

The approach in this paper can easily be extended to AAK type problems where k poles are allowed in the right half plane, with the proviso that one avoids the singular points (i.e., γ-optimal, the spectrum of the underlying Hankel operator, etc.). The change is that, instead of being outer, the inverse of the $(1, 1)$ block of the J-spectral factors is required to be in $\mathcal{RH}_\infty(k)$ (i.e., no more than k poles in the right half plane). The singular (optimal) case is, however, more involved, as a noncanonical factorization is required.

REFERENCES

[1] J. A. BALL AND N. COHEN, *Sensitivity minimization in an H^∞ norm: parametrization of all sub-optimal solutions*, Internat. J. Control, 46 (1987), pp. 785–816.

[2] J. A. BALL AND A. C. M. RAN, *Optimal Hankel norm model reductions and Weiner–Hopf factorization, I: The canonical case*, SIAM J. Control Optim., 25 (1987), pp. 362–383.

[3] ———, *Optimal Hankel norm model reductions and Wiener–Hopf factorization II: The noncanonical case*, Integral Equations and Operator Theory, 10 (1987), pp. 416–436.

[4] ———, *Hankel norm approximation for rational matrix functions in terms of realizations*, Conference on the Mathematical Theory of Networks and Systems, Stockholm, 1985, in Modeling, Identification, and Robust Control, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 285–296.

[5] M. BANKER, *Linear stationary quadratic games*, Proc. IEEE Conference on Decision and Control, (1973), pp. 193–197.

[6] H. BART, I. GOHBERG, AND M. A. KAASHOEK, *Minimal Factorization of Matrix and Operator Functions*, Birkhauser Verlag, Basel, 1979.

[7] D. S. BERNSTEIN AND W. M. HADDAD, LQG *control with an H_∞ performance bound: A Riccati equation approach*, IEEE Trans. Automat. Control, 34 (1989), pp. 293–305.

[8] J. C. DOYLE, *Lecture notes in advances in multivariable control*, ONR/Honeywell Workshop, Minneapolis, 1984.

[9] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. FRANCIS, *State-space solutions to standard H_2 and H_∞ control problems*, Proc. IEEE A.C.C., 1988; IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[10] B. A. FRANCIS, *A course in H_∞ control theory*, in Lecture notes in Control and Information Sciences 88, 2nd edition, Springer-Verlag, New York, 1987.

[11] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.

[12] K. GLOVER AND J. C. DOYLE, *State-space formulae for all stabilizing controllers that satisfy a H^∞ norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[13] K. GLOVER, D. J. N. LIMEBEER, J. DOYLE, E. M. KASENALLY, AND M. G. SAFONOV, *A characterization of all the solutions to the four block general distance problems*, SIAM J. Control Optim., to appear.

[14] J. W. HELTON, *Operator theory, analytic functions, matrices and electrical engineering* Conference Board of the Mathematical Sciences, Regional Conference Series in Mathematics 68, American Mathematical Society, Providence, RI, 1987.

[15] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, N.J., 1980.

[16] P. P. KHARGONEKAR, I. R. PETERSEN, AND M. A. ROTEA, *H_∞ optimal control with state feedback*, IEEE Trans. Automat. Control, 33 (1988), pp. 783–786.

[17] H. KIMURA AND R. KAWATANI, *Synthesis of H^∞ controllers based on conjugation*, Proc. of the 27th IEEE Conference on Decision and Control, Austin, Texas, 1988, pp. 7–13.

[18] D. J. N. LIMEBEER AND B. D. O. ANDERSON, *An interpolation theory approach to H^∞ controller degree bounds*, Linear Algebra Appl., 98 (1988), pp. 347–386.

[19] D. J. N. LIMEBEER, B. D. O. ANDERSON, P. P. KHARGONEKAR, AND M. GREEN, *A game theoretic approach to H_∞ control for time-varying systems*, submitted for publication.

[20] D. J. N. LIMEBEER AND G. D. HALIKIAS, *An analysis of pole zero cancellations in H^∞ optimal control problems of the second kind*, SIAM J. Control Optim., 26 (1988), pp. 646-677.

[21] D. J. N. LIMEBEER, E. M. KASENALLY, M. G. SAFONOV, AND I. JAIMOUKA, *A characterization of all the solutions to the four block general distance problem*, Proc. 27th IEEE Conference on Decision and Control, Austin, Texas, 1988, pp. 875-880.

[22] I. R. PETERSEN AND D. J. CLEMENTS, *J-spectral factorization and Riccati equations in problems of H^∞ optimization via state feedback*, preprint.

[23] M. G. SAFONOV, E. A. JONCKHEERE, M. VERMA, AND D. J. N. LIMEBEER, *Synthesis of positive real multivariable feedback systems*, Internat. J. Control, 45 (1987), pp. 817-842.

[24] M. G. SAFONOV AND D. J. N. LIMEBEER, *Simplifying the H^∞ theory via loop shifting*, Proc. IEEE Conference on Decision and Control, 1988, pp. 1399-1404.

[25] G. TADMOR, *H_∞ control in the time domain: the four block problem*, Math. Theory of Signals and Systems, to appear.

[26] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1372-1403, November 1990

006

DECOMPOSITION/COORDINATION ALGORITHMS IN STOCHASTIC OPTIMIZATION*

J.-C. CULIOLI† AND G. COHEN‡

Abstract. This paper considers an extension to the situation of stochastic programming of the Auxiliary Problem Principle formerly introduced in a deterministic setting to serve as a general framework for decomposition/coordination optimization algorithms. The idea is based upon that of the stochastic gradient, that is, independent noise realizations are considered successively along the iterations. As a consequence, deterministic subproblems are solved at each iteration whereas iterations fulfill the two tasks of coordination and stochastic approximation at the same time. Coupling cost function (expectation of some performance index) and *deterministic* coupling constraints are considered. Price (dual) decomposition (encompassing extensions of the Uzawa and Arrow-Hurwicz algorithms to this stochastic case) are studied as well as resource allocation (primal decomposition).

Key words. stochastic optimization, stochastic gradient, decomposition, coordination, dual methods, price decomposition, resource allocation, convergence of algorithms

AMS(MOS) subject classifications. 49D27, 49D29, 62L20, 65K10

1. Introduction.

For the optimization of large scale systems, the idea of decomposition has received much attention since the pioneering works of Dantzig and Wolfe [15] in linear programming, and Lasdon and his coauthors [21], [5] in convex programming. These were followed in the seventies by an abundant literature on the topic, starting with the book of Mesarovic, Macko, and Takahara [25]. Later on, an attempt was made to propose a unifying view of the field, first in the context of convex differentiable optimization (Cohen [6,], [7]), then in nondifferentiable optimization, (Cohen and Zhu [13]), and more recently for other variational problems (Cohen [11]) and games (Cohen [10]).

In this paper, we consider the case of stochastic optimization. Decomposition in the framework of stochastic optimization has already been considered in [3], [23], for example. But here the approach will be somewhat different. Let us explain this for the particular case of the following stochastic optimal control problem in discrete time:

$$\min \mathbb{E} \sum_{t=0}^{T} l_t(x_t, u_t, \omega_t) \tag{1}$$

$$x_{t+1} = f_t(x_t, u_t, \omega_t), \qquad t = 1, \cdots, T \tag{2}$$

where u is the control vector, x is the state vector, ω is a stochastic input, (1) is the objective function, (2) is the dynamics, and $\mathbb{E}$ denotes the mathematical expectation with respect to the probability law associated with the stochastic process (e.g., a white noise) $\{\omega_t\}$ and to the initial condition x_0.

Actually, for the problem to be well-posed, $\{u_t\}$ and $\{x_t\}$ must be specified as stochastic processes, which amounts to specifying the class of feedback laws that are

* Received by the editors January 16, 1989; accepted for publication (in revised form) October 17, 1989.
† Section Automatique, École des Mines de Paris, 35 Rue Saint-Honoré, 77305 Fontainebleau Cedex, France.

‡ Section Automatique, École des Mines de Paris, 35 Rue Saint-Honoré, 77305 Fontainebleau Cedex, France and the Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France.

allowed for the control. To fix ideas, suppose that only affine state feedbacks are allowed

$$u_t = C_t x_t + c_t \tag{3}$$

where C_t, respectively, c_t, is a time-dependent gain matrix, respectively, open-loop control vector, of appropriate dimensions. Let us further assume that the dynamics (2) are also affine and that the white noise ω_t is Gaussian. Then all stochastic processes will be Gaussian and they will be characterized by their first-order (vector) and second-order (matrix) moments. Lyapounov equations allow the calculations of the state covariance given the feedback gain sequence $\{C_t\}$ (see [23]).

If the system is made up of N interconnected systems with state and control vectors x^i, u^i, it may be required that C_t be block-diagonal with respect to these decompositions (decentralized feedback). However, unless the dynamics are completely decoupled, the covariance matrices will not be block-diagonal. Observe that a decomposition of those matrices and of the corresponding Lyapounov equations (say, by a relaxation technique as in [23]) would involve $N(N+1)/2$ blocks (instead of N blocks for vectors). It means that the interpretation of the potential subproblems, as problems of the same nature as the overall problem, would be lost. We refer the reader to [3] for a deeper discussion of this issue.

In what follows, we adopt a different point of view. We first note that in the field of deterministic optimization, decomposition/coordination algorithms and more classical optimization as the gradient algorithm are essentially of the same nature (that is of a variational nature). This is shown by the so-called *Auxiliary Problem Principle* which serves as a unifying framework [6], [7]. On the other hand, in the field of stochastic optimization, we may distinguish between global techniques such as dynamic programming and variational techniques such as the stochastic gradient [17], [20], which stems from the Robbins–Monro approximation technique [29]. It is therefore natural to combine the Auxiliary Problem Principle with the idea of the stochastic gradient algorithm which amounts essentially to considering successive independent realizations of the "noise," one at a time, and to performing successive gradient steps. Here, one gradient step will be replaced by the resolution of a *deterministic* auxiliary problem (corresponding to one particular realization of the noise), this auxiliary problem splitting up into N independent subproblems. Iterations will serve two purposes: coordination and approximation of the mathematical expectation. That is, at each iteration, coordination parameters will be updated and a new stochastic realization will be considered.

This point of view carries the same limitation as the stochastic gradient technique: because realizations are considered one at a time, it is out of the question to compute optimal "closed-loop" solutions; only "open-loop" solutions can be approximated. For example, in the case of the optimal stochastic control problem considered above, we can compute optimal "deterministic" values of parameters of an a priori feedback law such as (3), namely $\{C_t, c_t \mid t = 1, \cdots, T\}$, but not a general Markovian feedback $u_t = \varphi(x_t, t)$, as dynamic programming would allow.

Let us discuss another example: the problem of optimal investments in networks (see [2] for electrical networks and [27] for telecommunications networks). The general mathematical formulation is the following

$$\min_u \left[\alpha(u) + \mathbb{E}\left\{ \min_{v(u,\omega)} \beta(u, v(u,\omega), \omega) \text{ s.t. } \gamma(u, v(u,\omega), \omega) \leqq 0 \right\} \right]. \tag{4}$$

One must choose the capacities u of transmission lines of a network to minimize the sum of an investment cost α (increasing function) and of an optimal operation cost (decreasing with u). The latter cost is the expectation of the constrained minimum of

a stochastic cost β since the operation of the network involves perturbations ω (stochastic demand, failures, etc.) and on-line operational decisions v (routing, dispatching, etc.). These operational decisions are *closed-loop* variables (the operator has to react to observed situations) whereas the investment decisions are *open-loop* since one must decide upon the capacities of transmission lines at the beginning of a given planning period with only a statistical knowledge of the future situations.

In problem (1)-(2), as in problem (4), we may distinguish between closed-loop or stochastic variables (namely x and u—assuming that u is specified by (3)—in the former problem, v in the latter), and open-loop or deterministic variables (C and c in the former case, u in the latter). But we may also classify constraints as "almost sure" stochastic constraints[1] (involving ω and/or closed-loop variables) which generally represent physical laws (e.g., the dynamics (2)) and physical limitations (the inequality constraint in (4)), or classify them as deterministic constraints (involving only open-loop variables): the latter constraints would arise from the specification of admissible values for C and c in the former problem and for u in the latter. From the mathematical point of view, stochastic constraints involve stochastic or closed-loop Lagrange multipliers whereas deterministic constraints involve only determinisitc multipliers.

This classification delimits what is possible and what is not in terms of decomposition with our approach: essentially, any coupling arising through deterministic primal or dual variables (that is, in the latter case, through deterministic constraints) can be handled directly; on the other hand, if coordination is made necessary also because of some coupling arising from stochastic variables or constraints, then the corresponding coordination iterations cannot be "mixed" with those of stochastic approximation (which amounts to "visiting" independent noise realizations sequentially). In this latter case, it is of course always possible, for *fixed* values of all open-loop (primal and dual) variables and for some *fixed* realization of ω, to perform *all* the coordination iterations which are motivated by the coupling through closed-loop variables (if any), and *then* to update the open-loop variables and, at the same time, to draw a new independent noise realization. This means two iteration loops embedded one in the other.

The point of view that we just precisely described should be contrasted with another approach which consists in approximating the mathematical expectation by taking the average over a (relatively large) number of noise samples considered all together. This idea is generally exploited in the context of linear programming and can be found in the work of several authors among which we only quote [24] and [30] for the sake of brevity. The main advantage of this point of view over our approach is that closed-loop strategies can be handled (the closed-loop aspect is often referred to as "recourse" in this literature) but at the price of dealing with a very large scale problem. To try to remedy to this size increase, decomposition is sometimes considered but it mainly concerns the decomposition of the whole problem into subproblems corresponding to the individual noise samples, rather than to the decomposition of a physical system into interconnected subsystems. We do not insist more on this approach which is clearly in a quite different spirit from the idea of stochastic gradient.

Finally, in what follows two classes of problems will be discussed. Let $\mathcal{U}$ be a Hilbert space, U^f a closed convex (feasible) subset, $(\Omega, \mathfrak{A}, P)$ a probability space, and j a real-valued function over $\mathcal{U} \times \Omega$, lower semicontinuous and convex in u and measurable in ω. The first class of problems considered assumes the following general

[1] We will not deal here with constraints "in probability" but the same discussion applies as well to this case.

form:

$$\min_{u \in U^f} \int_\Omega j(u, \omega) P(d\omega) = \min_{u \in U^f} J(u). \tag{5}$$

It is assumed that the implicit constraint $u \in U^f$ is "simple" and will also appear as a constraint in the auxiliary problems. In particular, from the point of view of decomposition, it will be assumed decoupled, that is, if a decomposition $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_N$ is given, then

$$U^f = U_1^f \times \cdots \times U_N^f \tag{6}$$

where U_i^f is a closed convex subset of $\mathcal{U}_i$.

Notice that problem (5) may indeed represent a problem of type (4) if $j(u, \omega)$ is defined as

$$j(u, \omega) := \min_{v(u,\omega) \in V^f(u,\omega)} g(u, v(u, \omega), \omega). \tag{7}$$

However we recall the following two restrictions:

(i) the above minimization problem in v must be solved completely at each iteration before proceeding to the updating of u and to the drawing of a new value of ω;

(ii) the following condition must be satisfied, for, otherwise, u would be indirectly subject to stochastic constraints

$$\forall u \in U^f, \quad U^f(u, \omega) \neq \varnothing \text{ a.s.} \tag{8}$$

Because we assume that the constraint $u \in U^f$ induces no coupling, problem (5) represents a type of problem in which coupling arises only from the cost function. In order to deal with coupling arising from (deterministic!) constraints, we shall consider *explicit* constraints and we shall appeal to duality. Let Θ be an application from $\mathcal{U}$ to another Hilbert space $\mathcal{C}$, and C be a closed convex cone in $\mathcal{C}$. We consider the second class of problems

$$\min_{u \in U^f} \int_\Omega j(u, \omega) P(d\omega) \quad \text{s.t. } \Theta(u) \in -C. \tag{9}$$

For equality constraints, it suffices to set $C = \{0\}$. These equality or inequality constraints represent the only coupling constraints. We will consider two decomposition/coordination schemes, namely *price* or *dual* decomposition (see [21]) using ordinary or augmented Lagrangians, and *resource* or *right-hand side allocation* (see [5]).

As far as the former dual approach is concerned, apart from decomposition considerations, our general algorithms will encompass what may be considered the extensions of the Uzawa and Arrow-Hurwicz algorithms [1] to the case of stochastic optimization using either ordinary or augmented Lagrangians. This is not the first time that dual algorithms are considered in the framework of stochastic optimization: we are only aware of the work of Kushner and Clark [20] in this area.[2] We believe that our technique to study convergence, using convexity assumptions, is somewhat simpler and closer to the deterministic counterpart (see e.g., [13]).

We shall conclude this paper with some open problems and conjectures. This paper is based on the thesis dissertation of Culioli [14] a preliminary account of which was given in [12].

[2] See also the end of this Introduction.

After the first version of this paper was issued, a reviewer pointed out to us two main groups of works on numerical stochastic gradient (or "quasi-gradient") algorithms from the Russian literature. It would be rather lengthy to examine here the differences, weakness, or advantages of our work compared to those earlier works, as far as precise technical assumptions are concerned.[3] Therefore, we limit ourselves here to brief qualitative comments of this literature. It has in common with our approach the fact that it uses successive noise realizations one at a time.

The first group of algorithms uses the so-called "linearization" technique [16, pp. 215-218], [19, pp. 108-112]. At each stage of these algorithms, an auxiliary problem with a *linear* cost function has to be solved over a *bounded* feasible set. This cost function is built with an "average gradient" vector obtained by convex combination of all past information, be they stochastic gradients or "quasi-gradients" or even finite difference approximations. On the other hand, it will be clear that a basic common feature of all our algorithms is that the auxiliary problems that have to be solved at each stage are based on strongly convex (or convex-concave) auxiliary functions (plus linear correction terms), a feature that definitely precludes the use of, e.g., linear programming, to solve these auxiliary problems. This may sound like a limitation of our approach, but we expect a better numerical stability (i.e., convergence rate) as a counterpart to this effort of solving nonlinear (e.g., quadratic)—but generally decomposed—problems at each stage. However, we cannot support this claim by any numerical comparison experiments.

The second group of algorithms consists essentially of Arrow-Hurwicz-like algorithms in a stochastic context [28, pp. 116-126].[4] They are not quite different from what can be derived from our general algorithms studied in § 3. The main differences, apart from more or less restrictive technical assumptions, are again to be found in the technique of convergence proofs.

2. Coupling through the cost function only. In this section, we address the first class of problem (5). Let K be an auxiliary differentiable cost function (K' denotes the derivative) and $\{\varepsilon^k\}_{k\in\mathbb{N}}$ be an infinite sequence of numbers such that

$$\varepsilon^k > 0 \qquad \sum_{k=0}^{+\infty} \varepsilon^k = +\infty \qquad \sum_{k=0}^{+\infty} (\varepsilon^k)^2 < +\infty. \tag{10}$$

We propose the following algorithm.

ALGORITHM 1.

(i) *Pick up some* $u^0 \in U^f$; *set* $k=0$;

(ii) *At stage* k, *knowing* u^k, *draw an independent realization* ω^{k+1} *out of* Ω *according to the probability law* P *and a subgradient* r^k *out of* $\partial j(u^k, \omega^{k+1})$, *and compute* u^{k+1} *by solving*

$$\min_{u\in U^f} K(u) + \langle \varepsilon^k r^k - K'(u^k), u\rangle. \tag{11}$$

(iii) *Go back to* (ii) *with* $k \leftarrow k+1$.

As a first example, if $K(u) = \|u\|^2/2$ and if there is no constraint (i.e., $U^f = \mathcal{U}$), then (11) is readily solved and Algorithm 1 yields the stochastic (sub)gradient algorithm.

[3] Moreover, we had to read those papers in their original language since no English translations were provided, which makes very precise comments rather hazardous.

[4] We are also aware of an incomplete reference to S. P. Uryas'ev, *Arrow-Hurwicz algorithm with adaptively controlled step sizes* (1984), using rather sophisticated formulae for the step sizes.

With constraints ($u \in U^f \subset \mathcal{U}$), we get a projected subgradient formula $u^{k+1} = \Pi(u^k - \varepsilon^k r^k)$ where Π denotes the projection on U^f. From the decomposition point of view, assuming (6), it suffices to choose an *additive* auxiliary cost function $K(u) = \sum_i K_i(u_i)$ to realize that (11) splits up into N independent subproblems.

We can state the following convergence theorem.

THEOREM 1. (i) *We assume that $u \mapsto j(u, \omega)$ is convex, lower semicontinuous, subdifferentiable on U^f for all ω and that $\omega \mapsto j(u, \omega)$ is measurable on Ω, for all $u \in U^f$. If moreover J (defined in (5)) is coercive on u^f,*[5] *then (5) has solutions (the set of solutions is denoted U^* and any particular solution is denoted u^*).*

(ii) *We assume that K is differentiable and strongly convex with modulus $b > 0$ in U^f.*[6] *Then the solution u^{k+1} of (11) exists and is unique.*

(iii) *With assumption (10) and if j has linearly bounded subgradients (l.b.s. for short) in U^f, that is*

$$\exists c_1 > 0,\ c_2 > 0 : \forall u \in U^f, \quad \forall \omega, \quad \forall r \in \partial j(u, \omega),\ \|r\| \leqq c_1 \|u\| + c_2 \tag{12}$$

then

$$\lim_{k \to \infty} J(u^k) = J(u^*) \text{ a.s.}$$

the sequence $\{u^k\}$ is almost surely bounded and every cluster point (in the weak topology of $\mathcal{U}$) is a solution of (5).

(iv) *At last, if J is strongly convex, then U^* reduces to a singleton $\{u^*\}$ and $\{u^k\}$ almost surely strongly converges to u^*.*

The proofs of all theorems are gathered in an appendix.

Remark 1. One particular run of the algorithm corresponds to an infinite sequence $\{\omega^{k+1}\}_{k \in \mathbb{N}}$ of independent realizations in Ω. This is a trajectory of a stochastic process over the probability space $(\Omega, \mathfrak{A}, P)^{\otimes \mathbb{N}}$. One run also produces trajectories of other stochastic processes such as $\{u^k\}_{k \in \mathbb{N}}$ over the same probability space but with different state spaces. In the above theorem and in the rest of this paper, statements of "a.s. (almost sure) convergence" must be understood with respect to that probability space.

Note also that if $\mathfrak{F}^k$ denotes the sub-σ-algebra generated by $\omega^1, \cdots, \omega^k$ and if $\mathbb{E}^k$ denotes the conditional expectation knowing $\mathfrak{F}^k$, then for any function f

$$\mathbb{E}^k f(u^k) = f(u^k) \tag{13}$$

and

$$\mathbb{E}^k j(u^k, \omega^{k+1}) = \mathbb{E} j(u^k, \omega^{k+1}) = J(u^k). \tag{14}$$

Remark 2. Statement (iii) of Theorem 1 can be strengthened under an additional assumption which is met in particular if K is quadratic. Namely, if K' is continuous

---

[5] It means that for every sequence $\{u^k\} \subset U^f$ such that $\lim \|u^k\| = +\infty$, then $\lim J(u^k) = +\infty$.

[6] It means that

$$\exists b > 0 : \forall \alpha \in [0, 1], \quad \forall u, v \in U^f,\ K(\alpha u + (1-\alpha)v) \leqq \alpha K(u) + (1-\alpha) K(v) - \frac{b}{2} \alpha (1-\alpha) \|u - v\|^2.$$

This property is equivalent to the strong monotony of K', that is

$$\langle K'(u) - K'(v), u - v \rangle \geqq b \|u - v\|^2$$

or to the inequality

$$K(v) - K(u) \geqq \langle K'(u), v - u \rangle + \frac{b}{2} \|v - u\|^2.$$

from $\mathcal{U}$ equipped with the weak topology to $\mathcal{U}^*$ equipped with the weak-* topology, then it can be proved, using results of [8], that the whole sequence $\{u^k\}_{k\in\mathbb{N}}$ almost surely converges in the weak topology to some point in U^*.

The main idea of the Auxiliary Problem Principle is to locally replace a nonseparable cost function (here $j(u, \omega^{k+1})$) by a linear approximation (here $\langle r^k, u\rangle$) which is of course additive, and to reintroduce (strong) convexity through K which can, in addition, be chosen additive. Such a "linearization" is not necessary if some part of j is already additive or more generally separable. More specifically, suppose that

$$j(u, \omega) = g(u, \omega) + \Gamma\left(\sum_{i=1}^{N} h_i(u_i, \omega), \omega\right)$$

where g and $h = \sum_{i=1}^{N} h_i$ are convex functions of the same type as j previously, and $(x_1, \cdots, x_N) \mapsto \Gamma(x_1, \cdots, x_N, \omega)$ is a convex function from $\mathbb{R}^N$ to $\mathbb{R}$ which is nondecreasing with respect to every x_i for all other x_j and all ω (so that $\Gamma \circ h$ is also convex). For example, if Γ is simply the identity for all ω, j is the sum of g which is nonseparable and of h which is additive. The following variant of Algorithm 1 amounts to performing a "partial" linearization of j.

ALGORITHM 2. *In Algorithm* 1, *replace* (11) *by*

$$\min_{u\in U^f} K(u) + \langle \varepsilon^k r^k - K'(u^k), u\rangle + \langle \varepsilon^k \chi^k, h(u, \omega^{k+1})\rangle$$

where $\chi^k \in \partial\Gamma(x^k, \omega^{k+1})\big|_{x^k = h(u^k, \omega^{k+1})}$.

The proof of convergence of this variant can be found in [14]. The assumptions on g are the same as those on j in Theorem 1. Moreover Γ is assumed Lipschitz uniformly in ω, whereas for h it is assumed that

$$\begin{aligned} &\exists c_3 > 0,\ c_4 > 0:\ \forall u, v \in U^f, \forall \omega, \\ &\qquad |h(v, \omega) - h(u, \omega)| \leqq [c_3 \max(\|u\|, \|v\|) + c_4] \cdot \|v - u\|, \end{aligned} \tag{15}$$

which is equivalent to (12) if h is subdifferentiable. Note that

$$\max(\|u\|, \|v\|) \leqq \max(\|u\|, \|u\| + \|v - u\|) = \|u\| + \|v - u\|$$

so that (15) implies

$$|h(v, \omega) - h(u, \omega)| \leqq [c_3(\|u\| + \|v - u\|) + c_4] \cdot \|v - u\|. \tag{16}$$

3. Coupling through constraints and price decomposition. In this section, we consider problems of the form (9) where we recall that, in addition to J, only the explicit constraints involving Θ are intended to be coupling, that is, we still assume (6) as far as decomposition is concerned. Actually, to cover situations when the cost and the constraint functions are mixes of additive and nonadditive functions, we should consider the more general problem

$$\min_{u\in U^f} \int_\Omega (j(u, \omega) + g(u, \omega)) P(d\omega)^7 \quad \text{s.t. } \Theta(u) + \Xi(u) \in -C \tag{17}$$

where g and Ξ would be additive with respect to the decomposition of u whereas j and Θ would be at least subdifferentiable.

[7] Or else $\min_{u\in U^f} J(u) + G(u)$.

3.1. Some facts about duality. Recall that C is a closed convex cone in some Hilbert space $\mathscr{C}$. We say that a function Φ from $\mathscr{U}$ to $\mathscr{C}$ is *C-convex* if and only if

$$\forall \alpha \in [0,1], \forall u, v : \Phi(\alpha u + (1-\alpha)v) - \alpha\Phi(u) - (1-\alpha)\Phi(v) \in -C.$$

Note that if C reduces to $\{0\}$ (case of equality constraints), it means that Φ is affine. We say that Φ is *C-subdifferentiable* at u if and only if there exists a linear continuous operator φ from $\mathscr{U}$ to $\mathscr{C}$ such that

$$\forall v : \Phi(v) - \Phi(u) - \langle \varphi, v-u \rangle \in C.$$

The set of all such φ will be denoted $\partial\varphi(u)$. The conjugate cone C^* of C is defined by

$$C^* := \{ p \in \mathscr{C}^* \,|\, \langle p, c \rangle \geqq 0, \forall c \in C \}$$

where $\mathscr{C}^*$ denotes the topological dual of $\mathscr{C}$. Note that $C^* = \mathscr{C}^*$ if $C = \{0\}$. Moreover, if $p \in C^*$ and Φ is C-convex, then the functional $u \mapsto f_p(u) := \langle p, \Phi(u) \rangle$ is convex, and if $\varphi \in \partial\Phi(u)$, then $\varphi^T p \in \partial f_p(u)$, where φ^T is the adjoint operator of φ.

With a problem like

$$\min_{u \in U^f} F(u) \quad \text{s.t. } \Phi(u) \in -C \tag{18}$$

is associated a Lagrangian

$$L(u, p) := F(u) + \langle p, \Phi(u) \rangle$$

which has saddle points over $U^f \times C^*$ under convexity and other technical assumptions. The set of saddle points is of the form $U^* \times P^*$ and U^* is the set of solutions of (18). Classical algorithms to compute saddle points are the Uzawa and Arrow-Hurwicz algorithms [1]. The Uzawa algorithm consists of the following stages:

(i) $\min_{u \in U^f} F(u) + \langle p^k, \Phi(u) \rangle$ (yields u^{k+1});

(ii) update p^k by $p^{k+1} = \Pi(p^k + \rho\Phi(u^{k+1}))$ where ρ is a positive number and Π is the projection on C^*.

However, such an algorithm fails to converge to a solution of (18) if F is not strongly convex. As a matter of fact, this algorithm amounts to a "gradient" algorithm for maximizing the dual functional

$$\Psi(p) := \min_{u \in U^f} L(u, p). \tag{19}$$

This functional has Lipschitz gradients if F is *strongly* convex, and it is generally only subdifferentiable if F is not at least *strictly* convex.[8] In the latter case, even if p^k converges to an optimal p^*, it cannot be expected that the subgradients $\Phi(u^{k+1}) \in \partial\Psi(p^k)$ converge to an "optimal" value (say 0 in the equality constraint case). The situation improves if we use the *augmented* Lagrangian

$$L_c(u, p) := F(u) + \lambda_c(\Phi(u), p) \tag{20}$$

where c is a positive constant and

$$\lambda_c(t, p) := [\|\Pi(p + ct)\|^2 - \|p^2\|]/2c \tag{21}$$

is a convex-concave functional with derivatives

$$(\lambda_c)'_t(t, p) = \Pi(p + ct) \tag{22}$$

$$(\lambda_c)'_p(t, p) = [\Pi(p + ct) - p]/c. \tag{23}$$

[8] Observe that Ψ is always concave and we have that $\partial\Psi(p) = \overline{\text{co}}\, \Phi(\hat{U}(p))$ where $\overline{\text{co}}$ denotes the closure of the convex hull and $\hat{U}(p)$ denotes the set of optimal u in the minimization problem (19).

The fundamental reason is that, in this case, the corresponding Ψ_c (defined as Ψ but from L_c) is always Lipschitz differentiable. We refer the reader to [13] for details.

The basis of price decomposition [21] was the observation that the minimization stage (i) above splits into independent subproblems if and only if F and Φ are additive functions. This is no longer true if F or Φ is not additive or if L_c is used instead of L. These issues have been considered in [6], [7], [13]. We are going to extend this work to stochastic problems as (17) hereafter. But let us first notice that, for a problem as (9) for example (with $J \equiv 0$ and $\Theta \equiv 0$ to mimic a separable situation), a naive extension of the Uzawa algorithm as follows fails to work:

(i) $\min_{u \in U^f} g(u, \omega^{k+1}) + \langle p^k, \Xi(u) \rangle$ (yields u^{k+1});

(ii) update p^k by $p^{k+1} = \Pi(p^k + \varepsilon^k \Xi(u^{k+1}))$.

Here is a simple counterexample drawn from [14].

Example. Let $C = \{0\}$ (equality constraint), let $\Xi(u) = Du$ and $g(u, \omega) = \langle u, A(\omega)u \rangle / 2 + \langle b(\omega), u \rangle$. The optimality conditions for a pair (u^*, p^*) read

$$\bar{A}u^* + \bar{b} + D^T p^* = 0; \qquad Du^* = 0$$

where $\bar{A}$ is a shorter notation for $\mathbb{E}A(\omega)$ and likewise for $\bar{b}$. From these conditions, we get, assuming all inverses do exist,

$$p^* = -(D\bar{A}^{-1}D^T)^{-1} D\bar{A}^{-1}\bar{b}.$$

Now the above algorithm yields

(i) $A(\omega^{k+1})u^{k+1} + b(\omega^{k+1}) + D^T p^k = 0$

(ii) $p^{k+1} = p^k + \varepsilon^k Du^{k+1}$ which amounts to

$$(p^{k+1} - p^k)/\varepsilon^k = -D[A(\omega^{k+1})]^{-1} D^T p^k - D[A(\omega^{k+1})]^{-1} b(\omega^{k+1}).$$

Advocating the ordinary differential equation (ODE) technique of Ljung [22], it is seen that if any equilibrium point $\bar{p}$ exists for this algorithm, it must verify the equality

$$-D\bar{A}^{-1}D^T\bar{p} - D\mathbb{E}\{[A(\omega)]^{-1} b(\omega)\} = 0$$

showing that $\bar{p}$ has nothing to do with p^* in general.

3.2. The case of a strictly convex cost function and the use of ordinary Lagrangian.

3.2.1. Cost function not strongly convex. We follow a path similar to that of § 2 leading to Algorithm 1, except that we now deal with saddle points. To obtain our next algorithm, we choose an auxiliary function

$$T(u, p) = K(u) - \|p\|^2 / 2\gamma,$$

γ being a positive constant. We observe that the Lagrangian L relative to (17) is made up of a nonadditive term

$$M(u, p) = \int_\Omega m(u, p, \omega) P(d\omega) \quad \text{where } m(u, p, \omega) = j(u, \omega) + \langle p, \Theta(u) \rangle$$

and of an additive term

$$S(u, p) = \int_\Omega s(u, p, \omega) P(d\omega) \quad \text{where } s(u, p, \omega) = g(u, \omega) + \langle p, \Xi(u) \rangle$$

and we build the two successive auxiliary problems (assuming differentiability here, for the sake of simplicity)

(i) $\min_{u \in U^f} T(u, p^k) + \varepsilon^k s(u, p^k, \omega^{k+1}) + \langle \varepsilon^k m'_u(u^k, p^k, \omega^{k+1}) - T'_u(u^k, p^k), u \rangle$ (yields u^{k+1});

(ii) $\max_{p \in C^*} T(u^{k+1}, p) + \varepsilon^k s(u^{k+1}, p, \omega^{k+1}) + \langle \varepsilon^k m'_p(u^{k+1}, p^k, \omega^{k+1}) - T'_p(u^{k+1}, p^k), p \rangle$ (yields p^{k+1}).

In the more general case of nondifferentiable functions, this amounts to the following algorithm.

ALGORITHM 3.

(i) *Pick up some* $u^0 \in U^f$ *and* $p^0 \in C^*$; *set* $k=0$;

(ii) *At stage k, knowing* (u^k, p^k), *draw an independent realization* ω^{k+1} *out of* Ω *according to the probability law P, a subgradient* r^k *out of* $\partial j(u^k, \omega^{k+1})$, *a subgradient* θ^k *out of* $\partial\Theta(u^k)$ *and compute* u^{k+1} *by solving*

$$\min_{u \in U^f} K(u) + \varepsilon^k [g(u, \omega^{k+1}) + \langle p^k, \Xi(u)\rangle] + \langle \varepsilon^k r^k - K'(u^k), u\rangle + \varepsilon^k \langle (\theta^k)^T p^k, u\rangle. \tag{24}$$

(iii) *Update* p^k *by*

$$p^{k+1} = \Pi[p^k + \gamma\varepsilon^k(\Theta + \Xi)(u^{k+1})]. \tag{25}$$

(iv) *Go back to* (ii) *with* $k \leftarrow k+1$.

Two particular uses of this algorithm are of interest:

- When L is additive (i.e., $J \equiv 0$ and $\Theta \equiv 0$), which is the situation usually considered in price decomposition [21], with the simplest choice $K(u) = \|u\|^2/2$, we get a correct substitute to the naive extension of the Uzawa algorithm here mentioned above. The minimization stage now reads

$$\min_{u \in U^f} \|u - u^k\|^2/2 + \varepsilon^k [g(u, \omega^{k+1}) + \langle p^k, \Xi(u)\rangle].$$

- When, on the contrary, $G \equiv 0$ and $\Xi \equiv 0$, and with the same choice of K, we get the stochastic version of the Arrow–Hurwicz algorithm
  (i) $u^{k+1} = \pi[u^k - \varepsilon^k(r^k + (\theta^k)^T p^k)]$ where π denotes the projection on U^f
  (ii) $p^{k+1} = \Pi[p^k + \gamma\varepsilon^k \Theta(u^{k+1})]$.

THEOREM 2. (i) *We assume that* $j(\cdot, \omega)$ *and* $g(\cdot, \omega)$ *are convex, lower semicontinuous, that* $j(\cdot, \omega)$ *is subdifferentiable for all* ω *with* l.b.s. (*see* (12)), *that* $g(\cdot, \omega)$ *meets property* (15) *and that* $(j+g)(u, \cdot)$ *is measurable for all u. We assume that* Θ *and* Ξ *are C-convex and Lipschitz and that* Θ *is C-subdifferentiable. We assume that there exist saddle points* (u^*, p^*) *of L over* $U^f \times C^*$ (*this involves additional assumptions of coercivity of* $J+G$ *and of constraint qualification that we do not detail here*).

(ii) *We assume that K is differentiable and strongly convex with modulus* $b > 0$. *Then the solution* u^{k+1} *of* (24) *exists and is unique.*

(iii) *With assumption* (10), *we have that*

$$\lim_{k \to \infty} L(u^k, p^*) = L(u^*, p^*) \text{ a.s.}$$

for all $p^* \in P^*$ *and the sequence* $\{u^k\}$ *and* $\{p^k\}$ *are almost surely bounded. If* $J+G$ *is strictly convex, the sequence* $\{u^k\}$ *weakly converges to the unique solution* u^* *of* (17).

(iv) *At last, if* $J+G$ *is strongly convex, the convergence takes place in the strong topology.*

Remark 3. The assumption of strict convexity could be replaced by that of "stability in u of L" (see [13]), but practically the only case when it is easy to check for this assumption is that of strict convexity of the cost function.

Remark 4. It does not seem possible to get a convergence result for the dual sequence $\{p^k\}$ because, in some sense, such a convergence would be in connection with that of the subgradients $\{r^k\}$, which is out of the question.

3.2.2. Cost function strongly convex. In Theorem 2, we mentioned the case when $J+G$ is *strongly* convex. In this case, it can be proved that Ψ defined by (19) is not

only differentiable but that it has *Lipschitz derivatives* (the Lipschitz constant being ϕ^2/a when ϕ is the Lipschitz constant of $\Theta+\Xi$ and a is the strong convexity modulus of $J+G$—see footnote 6). Then the dual problem consists of maximizing this *smooth deterministic* function Ψ. To achieve such a task, it is enough to use "large steps" ρ (more generally, steps ρ^k that do not tend to zero) instead of "small steps" $\rho^k=\gamma\varepsilon^k$ as we did in Algorithm 3. This version is described in Algorithm 4 hereafter. However, due to the simultaneous use of small steps at the lower (primal) level and of large steps at the upper (dual) level, it is necessary to stabilize the algorithm by keeping the sequence $\{p^k\}$ in a bounded set (containing at least a dual solution p^*—see [13] for other instances of this kind).

To alleviate notations from now on, we come back to the simpler form (9) of problem (17) (i.e., we let $g\equiv 0$, $\Xi\equiv 0$).

ALGORITHM 4. *In Algorithm* 3, *let* $g\equiv 0$, $\Xi\equiv 0$ *and replace* (25) *by*

$$p^{k+1}=\Pi_\mu[p^k+\rho\Theta(u^{k+1})] \tag{26}$$

where ρ *is a positive number and* Π_μ *is the projection on the set* $B(0,\mu)\cap C^*$, $B(0,\mu)$ *being the closed ball with centre* 0 *and radius* $\mu>0$.

Remark 5. The projection Π_μ can be computed as $P_\mu\circ\Pi$ where P_μ denotes the projection on the ball $B(0,\mu)$ which is computed easily.

THEOREM 3. *We assume that* $B(0,\mu)$ *contains at least one optimal multiplier* p^*. *We strengthen the assumptions of Theorem* 2 *by assuming that* J *is strongly convex with modulus* a *and that the sequence* $\{\varepsilon^k\}$ *is nonincreasing. We call* τ *the Lipschitz constant of* Θ. *Then, if*

$$0<\rho<2a/\tau^2 \tag{27}$$

the sequence $\{u^k\}$ *generated by Algorithm* 4 *strongly converges to the unique solution* u^* *of* (9).

3.3. Cost function only convex and the use of augmented Lagrangian. As already discussed in § 3.1, the functional Ψ defined by (19) is generally nondifferentiable if $L(\cdot,p)$ is not at least strictly convex, whereas the functional Ψ_c defined as Ψ but after the augmented Lagrangian L_c (see (20)) has Lipschitz derivatives if the cost function and constraints are simply convex. We are thus going to reconsider Algorithm 3 or 4 with L_c instead of L.

This essentially amounts to considering algorithms involving the following two basic stages:

(i) Compute u^{k+1} by solving

$$\begin{aligned}\min_{u\in U^f} K(u)+\varepsilon^k g(u,\omega^{k+1})+\langle\varepsilon^k r^k-K'(u^k),u\rangle\\ +\varepsilon^k\langle(\lambda_c)'_t((\Theta+\Xi)(u^k),p^k),\theta^k\cdot u+\Xi(u)\rangle.\end{aligned} \tag{28}$$

(ii) Update p^k by

$$p^{k+1}=p^k+\gamma\varepsilon^k(\lambda_c)'_p((\Theta+\Xi)(u^{k+1}),p^k). \tag{29}$$

The expressions of $(\lambda_c)'_t$ and $(\lambda_c)'_p$ have been given in (22)–(23).

Remark 6. Note that there is no need to project the right-hand side of (29) on C^* because, with augmented Lagrangians, saddle points hold true on $U^f\times\mathscr{C}^*$ and not only on $U^f\times C^*$ as with ordinary Lagrangian.

From the point of view of decomposition, assuming (6), we observe that (28) splits up into independent subproblems provided that g and Ξ be additive functions

of u (otherwise, they should be subdifferentiable and they are incorporated into j and Θ, respectively) and that K be chosen additive, too.

A proof of convergence for the algorithm above was given in [14] (indeed, the part, g and Θ were not considered explicitly).

In the following, we shall consider a different version. This version follows a remark made by Mataoui[9] whose contribution is gratefully acknowledged. First, for the sake of simplicity, we drop again the "additive" terms ($g \equiv 0$, $\Xi \equiv 0$). Secondly, as we did in § 3.2.2, since Ψ_c is Lipschitz differentiable and deterministic (as was Ψ in that section), we are going to use "large steps" ρ instead of "small" steps $\gamma\varepsilon^k$ as above for the updating of p^k. Thirdly, as a consequence of mixing small steps at the lower level and large steps at the upper level, we need to keep p^k in a bounded set by projection on a ball $B(0, \mu)$ containing at least one optimal p^*.

ALGORITHM 5. *In Algorithm* 3, *replace* (24) *by*

$$\min_{u \in U^f} K(u) + \langle \varepsilon^k r^k - K'(u^k), u \rangle + \varepsilon^k \langle \Pi[p^k + c\Theta(u^k)], \theta^k \cdot u \rangle \tag{30}$$

and (25) *by*

$$p^{k+1} = P_\mu \left\{ p^k + \frac{\rho}{c} (\Pi[p^k + c\Theta(u^{k+1})] - p^k) \right\} \tag{31}$$

where P_μ has been defined in Remark 5.

Note that (30)-(31) are exactly (28)-(29) up to the differences explicitly stated above.

THEOREM 4. *We keep all the assumptions of Theorem* 2, *wherever relevant, and we assume, as in Theorem* 3, *that $\{\varepsilon^k\}$ is nonincreasing and that $B(0, \mu)$ contains at least one optimal p^*. We assume that*

$$0 < \rho < 2c. \tag{32}$$

Then the sequences $\{u^k\}$ and $\{p^k\}$ generated by Algorithm 5 *are almost surely bounded and every cluster point of $\{u^k\}$ (in the weak topology) is almost surely an optimal u^* for* (9).

4. Coupling through equality constraints and resource allocation.

4.1. The deterministic case. Let us first consider the deterministic constrained optimization problem

$$\min_{u \in \mathcal{U}} J(u) \quad \text{s.t. } Du = d. \tag{33}$$

Since we immediately consider *equality* constraints here, they have to be affine if we wish to remain in the framework of convex programming. Hence $D: \mathcal{U} \mapsto \mathcal{C}$ is a linear continuous operator and $d \in \mathcal{C}$.

Suppose that a decomposition of u is given. Since D is linear, it is additive, namely $Du = \sum D_i u_i$. Assume that the cost function J is also additive for simplicity. We may interpret the right-hand side d as a certain amount of resource that has to be shared among the N units. In price decomposition, the consumption $D_i u_i$ of each unit is priced with help of some "shadow" price p, and then each unit can minimize its own cost function $J_i(u_i)$ augmented by the cost of its individual consumption $\langle p, D_i u_i \rangle$. The role of coordination is to adjust p so that the total consumption Du balances the

[9] Section Automatique, École des Mines de Paris, Fontainebleau, France.

available resource amount d (which is mathematically possible if and only if there exists a saddle point for the Lagrangian associated with (33)).

The idea of resource (or right-hand side) allocation, also referred to as "feasible" or "primal" decomposition (see e.g., [18], [5]) is dual of that of price coordination. The total amount d is shared by coordination among the units according to some "feasible allocation" $v \in V^f$ where

$$V^f = \left\{ v \in \mathscr{C}^N \;\middle|\; \sum_{i=1}^{N} v_i = d \right\}. \tag{34}$$

Therefore each unit can optimize its own local behaviour by solving

$$\min_{u_i \in \mathscr{U}_i} J_i(u_i) \quad \text{s.t. } D_i u_i = v_i. \tag{35}$$

Assuming that these problems do have a solution for all considered allocations, the questions that should be clarified are the following:

- Does there exist an optimal allocation (that is, one making the corresponding "local" solutions of subproblems (35) optimal for (33))?
- How can such an optimal allocation be characterized in terms of information provided by the resolution of subproblems (35)?
- How to reach an optimal allocation computationally?

We refer the reader to [9, Part 1] for a detailed discussion of these issues. Let us mention briefly here that the answer to the first question is yes as long as there exists a solution u^* to the overall problem (33) (an optimal allocation is then given by $v_i^* = D_i u_i^*$ for all i). As for the second question, if there exist optimal Lagrange multipliers p_i^* associated with the allocation constraints in the subproblems (35), a *sufficient*[10] condition for some feasible allocation $v^* = (v_1^*, \cdots, v_N^*)$ to be optimal is that the corresponding multipliers verify $p_i^* = p_j^*$, for all i, j (which may be interpreted intuitively as the fact that the shared resource, at the amount at which it has been allocated to the local units, is marginally *equally* useful for each of them).

Finally, any coordination algorithm construction involves the following functional

$$\eta(v) = \sum_{i=1}^{N} \eta_i(v_i) \tag{36}$$

$$\eta_i(v_i) := \inf_{u_i \in \mathscr{U}_i} \sup_{p_i \in \mathscr{C}^*} J_i(u_i) + \langle p_i, D_i u_i - v_i \rangle$$

defined on V^f (say η takes the value $+\infty$ elsewhere).

Remark 7. Clearly, $\eta_i(v_i)$ is equal to the optimal cost value of (35) (equal to $+\infty$ when there is no feasible solution to (35)).

Obviously, solving (33) amounts to solving the coordination problem

$$\min_{v \in V^f} \eta(v). \tag{37}$$

One possible algorithm to solve (37) is the projected subgradient algorithm (not very often considered in the literature in this particular instance). It is well known that η is convex when (33) is a convex programming problem, and that, when a multiplier p (concatenation of the p_i's, optimal multipliers for the subproblems (35)) exists for some value of v, then $-p \in \partial\eta(v)$. The projected subgradient algorithm to solve (37) amounts to performing the following iterations:

[10] A necessary condition would involve at least uniqueness of the multipliers.

(i) For $i=1,\cdots,N$, solve (35) with v_i^k (yields u_i^{k+1} as primal solution and p_i^{k+1} as optimal multiplier);
(ii) Update v^k by $v^{k+1}=\bar{\omega}[v^k-\rho(-p^{k+1})]$ where ρ is a positive number and $\bar{\omega}$ is the projection on V^f.

Remark 8. The projection $\bar{\omega}$ is easily calculated. We have that

$$v_i^{k+1}=v_i^k+\rho p_i^{k+1}-\frac{1}{N}\left(\sum_{j=1}^{N}(v_j^k+\rho p_j^{k+1})-d\right)$$

$$=v_i^k+\rho\left(p_i^{k+1}-\frac{1}{N}\sum_{j=1}^{N}p_j^{k+1}\right)$$

since $v^k\in V^f$ after the first stage of the algorithm.

4.2. The stochastic case. Let us now come back to the stochastic case, that is, when $J(u)=\mathbb{E}j(u,\omega)$ in (33), but D and d are still deterministic (again, we are able to handle only deterministic constraints). Keeping our definition (36) of $\eta(v)$, the "master problem" (37) is still equivalent to the original problem (33). However, $\eta(v)$ is not simply the expectation of some cost function (indeed it is the inf sup of an expectation—see (36)) and the following algorithm is *not* a stochastic subgradient algorithm and there is *no way that it can work*:

(i) For $i=1,\cdots,N$, solve $\min_{u_i\in\mathcal{U}_i} j_i(u_i,\omega^{k+1})$ subject to $D_iu_i=v_i^k$ (yields u_i^{k+1} as primal solution and p_i^{k+1} as optimal multiplier);
(ii) Update v^k by $v^{k+1}=\bar{\omega}(v^k+\varepsilon^k p^{k+1})$.

We are going to derive a proper algorithm in a more systematic way, using the Auxiliary Problem technique, as already shown in § 3.2.1. But we first consider a slightly more general situation where we distinguish between an additive part g of the cost function and a nonadditive but subdifferentiable part j, as we did already in (17). Instead of (33), we thus consider

$$\min_{u\in\mathcal{U}}\mathbb{E}(j(u,\omega)+g(u,\omega))\quad\text{such that } Du=d \tag{38}$$

which is readily transformed into

$$\min_{v\in V^f}\min_{u\in\mathcal{U}}\sup_{p\in\mathcal{C}^*}\mathbb{E}(j(u,\omega)+g(u,\omega))+\langle p,Eu-v\rangle \tag{39}$$

where E is the block-diagonal operator

$$E:=\begin{pmatrix}D_1 & 0 & \cdots & 0\\ 0 & D_2 & & \vdots\\ \vdots & & \ddots & \vdots\\ 0 & \cdots & \cdots & D_N\end{pmatrix}.$$

We choose the auxiliary function[11]

$$T(u,v)=K(u)+\|v\|^2/2\gamma$$

where γ is a positive constant, and we share the expression in (39) into a "separable" term

$$s(u,p,\omega)=g(u,\omega)+\langle p,Eu\rangle$$

[11] In general, T should also be a function of p but here we choose it independent of p.

and its nonseparable complement

$$m(u, v, p, \omega) = j(u, \omega) - \langle p, v\rangle.$$

Remark 9. The reason that we consider the term $\langle p, Eu\rangle$ as a separable term whereas $\langle p, v\rangle$ is not considered so is that E is block-diagonal and u and p are going to be handled together at the subproblem level, whereas v is handled separately at the coordination level.

We then consider the following iterations (here in a differentiable case):

- $\min_{u\in\mathcal{U}} \sup_{p\in\mathcal{C}^*} T(u, v^k) + \varepsilon^k s(u, p, \omega^{k+1}) + \langle \varepsilon^k m'_u(u^k, v^k, p^k, \omega^{k+1}) - T'_u(u^k, v^k), u\rangle$ (yields (u^{k+1}, p^{k+1}));
- $\max_{v\in V^f} T(u^{k+1}, v) + \langle \varepsilon^k m'_v(u^{k+1}, v^k, p^{k+1}, \omega^{k+1}) - T'_v(u^{k+1}, v^k), v\rangle$ (yields v^{k+1}).

This finally yields the following algorithm (where we turn back the $\min_u \sup_p$ problem into a constrained minimization problem).

ALGORITHM 6.

(i) *Pick up some* $u^0 \in \mathcal{U}$ *and* $v^0 \in V^f$; *set* $k = 0$;

(ii) *At stage k, knowing* (u^k, v^k), *draw an independent realization* ω^{k+1} *out of* Ω *according to the probability law P, a subgradient* r^k *out of* $\partial j(u^k, \omega^{k+1})$, *and compute* (u^{k+1}, p^{k+1}) *by solving*

$$\min_{u\in\mathcal{U}} K(u) + \varepsilon^k g(u, \omega^{k+1}) + \langle \varepsilon^k r^k - K'(u^k), u\rangle \quad \text{s.t. } \varepsilon^k(Eu - v^k) = 0. \tag{40}$$

(iii) *Update* $v^k = (v_1^k, \cdots, v_N^k)$ *by*

$$v^{k+1} = \bar{\omega}(v^k + \gamma\varepsilon^k p^{k+1}) \tag{41}$$

or else, for $i = 1, \cdots, N$

$$v_i^{k+1} = v_i^k + \gamma\varepsilon^k\left(p_i^{k+1} - \frac{1}{N}\sum_{j=1}^{N} p_j^{k+1}\right).$$

(iv) *Go back to* (ii) *with* $k \leftarrow k+1$.

Remark 10. The factor ε^k multiplying the constraint appearing in (40) should not be dropped. Dropping it amounts to rescaling p^{k+1} and changing it into $\bar{p}^{k+1} = \varepsilon^k p^{k+1}$ (unless we also divide the whole cost function in (40) by the same factor ε^k). Such a rescaling does not in principle affect the dynamics of the algorithm, but numerically it does. As a matter of fact, $\bar{p}^{k+1}$ tends to zero with ε^k and the whole behaviour is modified because of the numerical "noise" (see [14] for numerical experiments).

If we assume, as in the deterministic case, that the cost function is additive, that is, $j \equiv 0$, and if we choose $K(u) = \|u\|^2/2$, (40) yields, for $i = 1, \cdots, N$

$$\min_{u_i\in\mathcal{U}_i} \frac{1}{2\varepsilon^k}\|u_i - u_i^k\|^2 + g_i(u_i, \omega^{k+1}) \quad \text{s.t. } D_i u_i = v_i^k$$

where, in application of Remark 10, we have divided both the cost function and the constraint by ε^k. This is the closest proper equivalent to (35) in the stochastic case.

THEOREM 5.

(i) *We assume that*:

— $j(\cdot, \omega)$ *and* $g(\cdot, \omega)$ *are convex, lower semicontinuous*;

— $j(\cdot, \omega)$ *is subdifferentiable for all* ω *with* l.b.s. (*see* (12)), $g(\cdot, \omega)$ *meets property* (15), $(j+g)(u, \cdot)$ *is measurable for all u, and* $J + G$ *is coercive on* $\mathcal{U}$;

— D *is linear and continuous and* $d \in \operatorname{Int}(\operatorname{Im} D)$.

Then the Lagrangian associated with (38) *has a saddle-point* (u^*, p^*).

(ii) *We assume that*:
- — *K is differentiable and strongly convex with modulus $b>0$, and K' is Lipschitz with constant B;*
- — *Each D_i is linear, continuous and surjective onto $\mathscr{C}$.*

Then the Lagrangian associated with (40) *has a saddle point (u^{k+1}, p^{k+1}) and u^{k+1} is unique.*

(iii) *For γ small enough,*[12] *and if*

$$\exists \zeta : \forall k \in \mathbb{N}, \quad \varepsilon^k \leqq \zeta \varepsilon^{k+1} \tag{42}$$

then the sequences $\{u^k\}, \{p^k\}, \{v^k\}$ generated by Algorithm 6 are almost surely bounded and every cluster point of $\{u^k\}$ in the weak topology is almost surely a solution u^* of (38).

(iv) *Finally the convergence takes place in the strong topology towards the unique u^* if and only if $J+G$ is strongly convex.*

Remark 11. Admittedly, the assumption that each D_i (or E) is onto is rather strong, but we do not know how to alleviate it. It implies that Problem (40) is well defined for all v^k. It is also equivalent to the fact that each $D_iD_i^T$ (or EE^T) is strongly monotone, which is the property used in the proof.

Remark 12. The assumption (42) is mild. It holds true, for example, if $\{\varepsilon^k/\varepsilon^{k+1}\}$ is nonincreasing, which is the case for the usual sequences one uses to meet (10).

5. Conclusion. In this paper, we have considered stochastic convex programming problems where the cost function is the expectation of some performance index corrupted by noise and where the constraints, if any, are deterministic. We were interested in decomposition algorithms, but the algorithms presented are of interest even without this feature in mind.

We attempted to extend the so-called Auxiliary Problem Principle, previously introduced in a deterministic setting to provide a general framework for decomposition algorithms, to that situation of stochastic optimization. However, we wanted to preserve the idea of stochastic gradient, which amounts to considering a single independent realization of the noise at each iteration. In this way, the auxiliary (decomposed) problem to be solved at each stage is deterministic. The iterations serve two purposes at the same time: they coordinate the subproblem solutions to make them converge to the overall optimum—this is the coordination task—and they visit many independent random realizations—this is the stochastic approximation task according to the scheme first introduced by Robbins and Monro [29].

However, this approach bears its own limitation, namely that only "open-loop" or deterministic variables can be approximated in this way. This as well applies to dual variables with the consequence that stochastic coupling constraints cannot be handled by coordination (unless coordination iterations be pushed to their end while keeping the noise realization fixed). This is a serious limitation since, in problem (1)–(2) for example, the dynamics should not be coupling.

There is however a possible way of dealing with stochastic constraints, be they coupling or not. The idea is to avoid manipulating multipliers by appealing to some kind of penalty technique. This is offered as a conjecture hereafter. Let us first consider a stochastic problem with almost sure constraints but with a *finite set* Ω. Say, $\Omega = \{\omega_1, \cdots, \omega_m\}$ and π_i is the probability associated with realization ω_i. The problem

[12] See Remark 13 in § E of the Appendix to see how small γ should be.

can be formulated as follows:

$$\min_u \sum_{i=1}^{m} \pi_i j(u, \omega_i) \quad \text{s.t. } \Theta(u, \omega_i) \in -C, \quad \forall i. \tag{43}$$

Let u^* be a solution. It is known [31] that $c \in -C \Leftrightarrow \Pi(c) = 0$ (recall that Π is the projection over C^*) so that the constraints can be written again as equality constraints. With each constraint, a multiplier p_i can be associated and, assuming that optimal multipliers p_i^* do exist, it is known (under the name of "exact penalty" technique [4]) that if Q is large enough (compared to $\sup_i \|p_i^*\|$), then u^* can be found by solving

$$\min_u \sum_{i=1}^{m} [\pi_i j(u, \omega_i) + Q\|\Pi(\Theta(u, \omega_i))\|].$$

A simple manipulation shows that u^* can also be found by solving

$$\min_u \mathbb{E}[j(u, \omega) + Q'\|\Pi(\Theta(u, \omega))\|] \tag{44}$$

with Q' larger than $Q/\inf_i \pi_i$ (indeed Q' larger than $\sup_i \|p_i^*\|/\pi_i$ would be enough). This problem is of type (5) and can be solved by the algorithms of § 2.

We see that if Ω now becomes infinite, the probabilities π_i will approach zero, and, unless the optimal multipliers associated with constraints corresponding to weak probabilities are small (and there is no reason why this should be the case), Q' will approach $+\infty$. However, in a practical problem, it is reasonable to withdraw realizations with very small probabilities from the problem formulation before considering almost sure constraints. Or else, one should consider only constraints in probability (say, the constraint must be met with probability 0.95).

It is conjectured that the right theorem would be: "*The solution of* (44) *is a solution of a version of* (43) *with constraints in probability and this probability tends to* 1 *when* $Q' \to +\infty$." If some sort of theorem of this kind is true, then the solution of problems with stochastic constraints can be approximated by problems that we know how to solve and decompose by the algorithms studied in this paper. However, this is not always satisfactory (think of dealing with the constraints (2) by such a technique).

Finally, comparing our work with the literature we are aware of, although the stochastic gradient algorithm and its variants have been largely considered, algorithms appealing to duality to handle (deterministic) constraints do not seem to have been thoroughly studied. One noticeable exception is the work by Kushner and Clark [20]. Although a direct comparison of their results with ours is not straightforward since they do not always study the same algorithms,[13] it seems that our assumptions are often less restrictive (no differentiability, only subdifferentiability required, no strong nor even strict convexity required except for Algorithms 3 and 4 using ordinary Lagrangians) and, above all, our technique of proof seems simpler than theirs. See also the discussion relative to [28] at the end of the introduction.

As for the resource allocation algorithm (§ 4.2), it does not seem to have been considered elsewhere in a stochastic context.

Appendix: proofs of convergence theorems.

Two technical lemmas from [13]. For the sake of completeness, we quote here two technical lemmas drawn from [13] that will be repeatedly referred to in this appendix.

[13] They study a penalty algorithm that we do not consider but they do not study the augmented Lagrangian algorithm; they very often appeal to projections on bounded sets, what we do only when mixing "large" and "small steps."

LEMMA 4 of [13]. *Let f be a Lipschitz functional on a Hilbert space $\mathcal{U}$ and consider the sequences $\{u^k\}_{k\in\mathbb{N}}\subset\mathcal{U}$ and $\{\varepsilon^k\}_{k\in\mathbb{N}}\subset\mathbb{R}^+$ such that*

$$\exists\delta:\forall k\in\mathbb{N},\quad \|u^{k+1}-u^k\|\leqq\delta\varepsilon^k$$

$$\sum_{k\in\mathbb{N}}\varepsilon^k=+\infty$$

$$\exists\mu:\sum_{k\in\mathbb{N}}\varepsilon^k|f(u^k)-\mu|<+\infty.$$

Then $\lim_{k\to+\infty}f(u^k)=\mu$.

LEMMA 5 of [13]. *Let $\{x^k\}_{k\in\mathbb{N}}\subset\mathbb{R}^+$ and $\{\alpha^k\}_{k\in\mathbb{N}}\subset\mathbb{R}^+$ such that*

$$\sum_{k\in\mathbb{N}}\alpha^k<+\infty.$$

Let X^k denote $\sup_{\ell\leqq k}x^\ell$, *and assume that*

$$x^k\leqq\sum_{\ell=1}^{k-1}a^\ell X^{\ell+1}+\delta^k$$

and that $\delta^k\leqq\delta$, for all $k\in\mathbb{N}$. Then the sequence $\{x^k\}_{k\in\mathbb{N}}$ is bounded.

A. Proof of Theorem 1. The proof of statements (i) and (ii) is based on classical arguments and we skip it here. The solution u^{k+1} of (11) is characterized by the following variational inequality:

$$\forall u\in U^f,\quad \langle K'(u^{k+1})-K'(u^k)+\varepsilon^k r^k,\,u-u^{k+1}\rangle\geqq 0. \tag{45}$$

For some solution u^* of (5), we consider the Lyapounov function Λ

$$\Lambda(u)=K(u^*)-K(u)-\langle K'(u),\,u^*-u\rangle.$$

Note that

$$\Lambda(u)\geqq b\|u-u^*\|^2/2 \tag{46}$$

from the strong convexity of K.

From (12), and using $(a+b)^2\leqq 2(a^2+b^2)$, it should be clear that there exist positive constants c_5 and c_6 such that

$$\begin{aligned}\|r^k\|^2&\leqq c_5\|u^k-u^*\|^2+c_6\\&\leqq 2c_5\Lambda(u^k)/b+c_6,\end{aligned} \tag{47}$$

the latter inequality arising from (46).

On the other hand, we have that

$$\begin{aligned}b\|u^{k+1}-u^k\|^2&\leqq\langle K'(u^{k+1})-K'(u^k),\,u^{k+1}-u^k\rangle && \text{(strong monotony)}\\&\leqq\varepsilon^k\langle r^k,\,u^k-u^{k+1}\rangle && \text{(from (45) with } u=u^k)\\&\leqq\varepsilon^k\|r^k\|\cdot\|u^{k+1}-u^k\| && \text{(Schwarz inequality)}\end{aligned}$$

hence

$$\|u^{k+1}-u^k\|\leqq\varepsilon^k\|r^k\|/b. \tag{48}$$

It follows that

$$\begin{aligned}\varepsilon^k\langle r^k,\,u^k-u^{k+1}\rangle&\leqq\varepsilon^k\|r^k\|\cdot\|u^k-u^{k+1}\| && \text{(Schwarz inequality)}\\&\leqq(\varepsilon^k)^2\|r^k\|^2/b && \text{(from (48))}\end{aligned}$$

hence

$$\varepsilon^k\langle r^k, u^k - u^{k+1}\rangle \leqq (\varepsilon^k)^2(c_7\Lambda(u^k)+c_8) \tag{49}$$

from (47) and for some positive constants c_7 and c_8.

We now study the variation of Λ over one stage of the algorithm, namely $\Lambda(u^{k+1}) - \Lambda(u^k)$ denoted by $\delta_k^{k+1}\Lambda$ for short. From the definition of Λ, we get

$$\delta_k^{k+1}\Lambda = \underbrace{K(u^k) - K(u^{k+1}) - \langle K'(u^k), u^k - u^{k+1}\rangle}_{A_1} + \underbrace{\langle K'(u^k) - K'(u^{k+1}), u^* - u^{k+1}\rangle}_{A_2}.$$

Using (45) with $u = u^*$, we see that

$$A_2 \leqq \varepsilon^k\langle r^k, u^* - u^{k+1}\rangle = \underbrace{\varepsilon^k\langle r^k, u^* - u^k\rangle}_{B_1} + \underbrace{\varepsilon^k\langle r^k, u^k - u^{k+1}\rangle}_{B_2}$$

and

$$B_1 \leqq \varepsilon^k(j(u^*, \omega^{k+1}) - j(u^k, \omega^{k+1}))$$

from the convexity of $j(\cdot, \omega^{k+1})$ and the definition of r^k, whereas B_2 is bounded by (49). Also, from the convexity of K, A_1 is nonpositive.

Collecting everything, we get

$$\delta_k^{k+1}\Lambda \leqq \varepsilon^k(j(u^*, \omega^{k+1}) - j(u^k, \omega^{k+1})) + (\varepsilon^k)^2(c_7\Lambda(u^k)+c_8).$$

Taking the conditional expectation with respect to $\mathcal{F}^k$ and remembering (13)-(14), we get

$$\mathbb{E}^k\Lambda(u^{k+1}) - \Lambda(u^k) \leqq \varepsilon^k(J(u^*) - J(u^k)) + (\varepsilon^k)^2(c_7\Lambda(u^k)+c_8). \tag{50}$$

Let $y^k := \mathbb{E}\Lambda(u^k)$ and notice that $J(u^*) \leqq J(u^k)$ by definition of u^*. Taking the expectation in (50), we get

$$y^{k+1} - y^k \leqq \alpha^k y^k + \beta^k \tag{51}$$

where $\{\alpha^k\}$ and $\{\beta^k\}$ are convergent series of positive numbers. Using Lemma 5 of [13], we conclude that the sequence $\{y^k\}$ is bounded.

Coming back to (50), and summing up for all values of k, we get

$$\sum_{k\in\mathbb{N}} \mathbb{E}(\mathbb{E}^k\Lambda(u^{k+1}) - \Lambda(u^k))^+ \leqq \sum_{k\in\mathbb{N}} (\alpha^k y^k + \beta^k) < \infty$$

where $(x)^+$ denotes $\max(x, 0)$. This, together with $\inf_{k\in\mathbb{N}} \mathbb{E}(\Lambda(u^k)) > -\infty$ (which is obvious since Λ is nonnegative—see (46)), implies that $\{\Lambda(u^k)\}$ is a quasimartingale and therefore that it almost surely converges to some random variable with finite expectation [26, pp. 49-51]. Thus it is almost surely bounded and so are the sequences $\{r^k\}$ from (47) and $\{u^k\}$ from (46). Therefore $\{u^k\}$ is almost surely weakly compact and has almost surely cluster points in the weak topology. We now prove that any such cluster point, say $\bar{u}$, is indeed a solution u^* of (5).

For this, we first need to prove that $J(u^k)$ almost surely converges to $J(u^*)$. Taking again the expectation in (50) and summing up, we have that

$$\begin{aligned}\sum_{k\in\mathbb{N}} \varepsilon^k\mathbb{E}(J(u^k) - J(u^*)) &\leqq \sum_{k\in\mathbb{N}} (y^k(1+\alpha^k) - y^{k+1} + \beta^k) \\ &\leqq \sum_{k\in\mathbb{N}} (M\alpha^k + \beta^k) + y^0 \\ &< \infty\end{aligned} \tag{52}$$

since y^k is positive and bounded by some M. Each term $J(u^k)-J(u^*)$ being nonnegative, this proves that, almost surely,

$$\sum_{k\in\mathbb{N}} \varepsilon^k (J(u^k)-J(u^*)) < \infty. \tag{53}$$

It is easy to see that J is Lipschitz over any set on which its subgradients are bounded. We also have that $\|u^{k+1}-u^k\| \leqq c_9 \varepsilon^k$ from (48) and for some positive constant c_9. Because of these facts and with (53), we conclude that $J(u^k)$ almost surely converges to $J(u^*)$ using Lemma 4 of [13].

Now if we consider a cluster point $\bar{u}$ of the sequence $\{u^k\}$ in the weak topology, $\bar{u}$ belongs to U^f which is closed and convex, hence weakly closed. Since J is convex and lower semicontinuous, hence weakly lower semicontinuous, we have, for a subsequence $\{u^{k_i}\}$ weakly converging to $\bar{u}$

$$J(\bar{u}) \leqq \liminf_{k_i\to\infty} J(u^{k_i}) = J(u^*)$$

proving that indeed $\bar{u}$ is equal to some u^* almost surely.

We complete the proof by considering the case when J is strongly convex with modulus a, in which case u^* is unique. This solution is characterized by the variational inequality

$$\exists R^* \in \partial J(u^*): \forall u \in U^f, \qquad \langle R^*, u-u^*\rangle \geqq 0.$$

Then

$$\begin{aligned} J(u^k)-J(u^*) &\geqq \langle R^*, u^k-u^*\rangle + a\|u^k-u^*\|^2/2 \\ &\geqq a\|u^k-u^*\|^2/2, \end{aligned}$$

but since $J(u^k)$ converges to $J(u^*)$ almost surely, it follows that $\|u^k-u^*\|$ tends to zero almost surely.

B. Proof of Theorem 2. The solution u^{k+1} of (24) is characterized by the following variational inequality:

$$\begin{aligned} &\forall u \in U^f, \\ &\langle K'(u^{k+1})-K'(u^k)+\varepsilon^k r^k, u-u^{k+1}\rangle + \varepsilon^k (g(u,\omega^{k+1})-g(u^{k+1},\omega^{k+1})) \\ &\quad + \varepsilon^k \langle p^k, \theta^k\cdot(u-u^{k+1})\rangle + \varepsilon^k \langle p^k, \Xi(u)-\Xi(u^{k+1})\rangle \geqq 0. \end{aligned} \tag{54}$$

We have that

$$\begin{aligned} b\|u^{k+1}-u^k\|^2 &\leqq \langle K'(u^{k+1})-K'(u^k), u^{k+1}-u^k\rangle && \text{(strong monotony)} \\ &\leqq \varepsilon^k [\langle r^k, u^k-u^{k+1}\rangle + g(u^k,\omega^{k+1})-g(u^{k+1},\omega^{k+1}) \\ &\quad + \langle p^k, \theta^k\cdot(u^k-u^{k+1})+\Xi(u^k)-\Xi(u^{k+1})\rangle] && \text{(from (54) with } u=u^k) \\ &\leqq \varepsilon^k [\|r^k\| + c_5(\|u^k\|+\|u^{k+1}-u^k\|)+c_6+\|p^k\|(\tau+\xi)]\cdot\|u^{k+1}-u^k\| \end{aligned}$$

where, in the last inequality, we have used the Schwarz inequality, property (16) for g, the Lipschitz constants τ and ξ of Θ and Ξ, respectively, and the fact that for all $\theta\in\partial\Theta(u)$, $\|\theta\|\leqq\tau$. It follows that

$$b\|u^{k+1}-u^k\| \leqq \varepsilon^k [\|r^k\| + c_5(\|u^k\|+\|u^{k+1}-u^k\|)+c_6+\|p^k\|(\tau+\xi)].$$

Since ε^k goes to zero, $c_5\varepsilon^k \leqq b/2$ for k large enough, and we may assume that this is

already true for $k=0$ without loss of generality. Hence

$$\begin{aligned}\|u^{k+1}-u^k\| &\leqq \varepsilon^k[\|r^k\|+c_7\|u^k\|+c_8+\|p^k\|(\tau+\xi)] \\ &\leqq \varepsilon^k[c_9\|u^k\|+c_{10}+\|p^k\|(\tau+\xi)],\end{aligned} \tag{55}$$

the latter inequality arising from (12).

For some saddle point (u^*, p^*), consider now the Lyapounov function

$$\Lambda(u,p)=K(u^*)-K(u)-\langle K'(u), u^*-u\rangle+\|p-p^*\|^2/2\gamma$$

and observe that

$$\|p^k-p^*\|^2\leqq 2\gamma\Lambda(u^k,p^k) \quad \text{and} \quad \|u^k-u^*\|^2\leqq 2\Lambda(u^k,p^k)/b. \tag{56}$$

We study the difference $\Lambda(u^{k+1},p^{k+1})-\Lambda(u^k,p^k)$ (denoted by $\delta_k^{k+1}\Lambda$ again)

$$\begin{aligned}\delta_k^{k+1}\Lambda = &\underbrace{K(u^k)-K(u^{k+1})-\langle K'(u^k), u^k-u^{k+1}\rangle}_{A_1} \\ &+\underbrace{\langle K'(u^k)-K'(u^{k+1}), u^*-u^{k+1}\rangle}_{A_2} \\ &+\underbrace{(\|p^{k+1}-p^*\|^2-\|p^k-p^*\|^2)/2\gamma}_{A_3}.\end{aligned}$$

We have that

$$A_1\leqq -b\|u^{k+1}-u^k\|^2/2 \tag{57}$$

from the last inequality in Footnote 6. As for A_2, we consider (54) with $u=u^*$. This yields

$$\begin{aligned}A_2\leqq \varepsilon^k[&\langle r^k, u^*-u^{k+1}\rangle+g(u^*,\omega^{k+1})-g(u^{k+1},\omega^{k+1}) \\ &+\langle p^k, \theta^k\cdot(u^*-u^{k+1})+\Xi(u^*)-\Xi(u^{k+1})\rangle].\end{aligned} \tag{58}$$

Using the convexity of j and Θ together with the definition of r^k and θ^k, the Schwarz inequality, property (16) for g and the Lipschitz property of Θ (which implies also that $\theta^k\leqq\tau$), we get

$$\begin{aligned}A_2\leqq \varepsilon^k[&(j+g)(u^*,\omega^{k+1})-(j+g)(u^k,\omega^{k+1}) \\ &+(\|r^k\|+c_5(\|u^k\|+\|u^{k+1}-u^k\|)+c_6+\tau\|p^k\|)\cdot\|u^{k+1}-u^k\| \\ &+\langle p^k, \Theta(u^*)-\Theta(u^k)\rangle+\langle p^k, \Xi(u^*)-\Xi(u^{k+1})\rangle].\end{aligned} \tag{59}$$

Finally for A_3, we use (25) together with a similar equality for p^* (which is equivalent to the left-hand side inequality of the saddle point), namely

$$p^*=\Pi[p^*+\gamma\varepsilon^k(\Theta+\Xi)(u^*)]. \tag{60}$$

Because the projection does not expand distances, we get

$$\|p^{k+1}-p^*\|\leqq\|p^k-p^*+\gamma\varepsilon^k[(\Theta+\Xi)(u^{k+1})-(\Theta+\Xi)(u^*)]\|,$$

taking the square

$$A_3\leqq\varepsilon^k\langle p^k-p^*, (\Theta+\Xi)(u^{k+1})-(\Theta+\Xi)(u^*)\rangle+\gamma\phi^2(\varepsilon^k)^2\|u^{k+1}-u^*\|^2/2$$

where we have used the Lipschitz constant ϕ of $\Theta+\Xi$ (which is less than $\tau+\xi$).

Summing up these inequalities for A_1, A_2, A_3, we get

$$\begin{aligned}\delta_k^{k+1}\Lambda \leqq\ & \varepsilon^k[\ell(u^*, p^*, \omega^{k+1}) - \ell(u^k, p^*, \omega^{k+1})] \\ & + \underbrace{\varepsilon^k[\langle p^k, \Theta(u^{k+1}) - \Theta(u^k)\rangle + \langle p^*, (\Theta+\Xi)(u^k) - (\Theta+\Xi)(u^{k+1})\rangle]}_{B_1} \\ & + \underbrace{\varepsilon^k[\|r^k\| + c_5\|u^k\| + c_6 + \tau\|p^k\|] \cdot \|u^{k+1} - u^k\|}_{B_2} \\ & + \underbrace{\gamma\phi^2(\varepsilon^k)^2\|u^{k+1} - u^*\|^2/2 + (\varepsilon^k c_5 - b/2)\|u^{k+1} - u^k\|^2}_{B_3}\end{aligned}$$

where we have set

$$\ell(u, p, \omega) := (j+g)(u, \omega) + \langle p, (\Theta+\Xi)(u)\rangle. \tag{61}$$

The last terms above can be bounded as follows. The Schwarz inequality and the Lipschitz property of Θ and Ξ yields a bound of B_1 that can be incorporated into B_2 using different constants. Also, remembering (12) and (55), and using standard manipulations, we get

$$\begin{aligned}B_1 + B_2 &\leqq (\varepsilon^k)^2[c_{11}\|u^k - u^*\|^2 + c_{12}\|p^k - p^*\|^2 + c_{13}] \\ &\leqq (\varepsilon^k)^2[c_{14}\Lambda(u^k, p^k) + c_{13}]\end{aligned} \tag{62}$$

the latter from (56). On the other hand

$$B_3 \leqq \gamma\phi^2(\varepsilon^k)^2\|u^k - u^*\|^2 + (\gamma\phi^2(\varepsilon^k)^2 + c_5\varepsilon^k - b/2)\|u^{k+1} - u^k\|^2.$$

The former term in the right-hand side can be bounded from above by an expression similar to (62) using (56) again, whereas the latter term is nonnegative for k large enough (and we may assume that this is already true for $k = 0$ without loss of generality).

Collecting everything, we get the basic inequality

$$\delta_k^{k+1}\Lambda \leqq \varepsilon^k(\ell(u^*, p^*, \omega^{k+1}) - \ell(u^k, p^*, \omega^{k+1})) + (\varepsilon^k)^2(c_{15}\Lambda(u^k, p^k) + c_{16}). \tag{63}$$

This inequality plays the part of (50) in the previous proof. Thus, proceeding exactly as previously, we can prove the following results:

- $\{\Lambda(u^k, p^k)\}$ is a quasimartingale, which converges almost surely to some random variable with finite expectation;
- the sequences $\{u^k\}$, $\{r^k\}$, $\{p^k\}$ are almost surely bounded;
- the sequence $\{L(u^k, p^*) = \mathbb{E}^k\ell(u^k, p^*, \omega^{k+1})\}$ almost surely converges to $L(u^*, p^*)$;
- therefore, using the (weak) lower-semicontinuity of $L(\cdot, p^*)$, each cluster point of $\{u^k\}$ in the weak topology minimizes $L(\cdot, p^*)$; but since $(J+G)$ is assumed to be *strictly* convex, u^* is the only such minimizer and the whole sequence $\{u^k\}$ weakly converges to the unique u^*.

The case when $J+G$ is *strongly* convex is handled exactly as it was previously.

C. Proof of Theorem 3. We indicate the main modifications that must be applied to the previous proof. The calculations leading to (55) are still valid. The Lyapounov function to be considered now is

$$\Lambda^k(u, p) = K(u^*) - K(u) - \langle K'(u), u^* - u\rangle + (\varepsilon^k/2\rho)\|p - p^*\|^2. \tag{64}$$

The inequalities in (56) now read

$$\|p^k\| \leqq \mu \quad \text{and} \quad \|u^k - u^*\|^2 \leqq 2\Lambda^k(u^k, p^k)/b; \tag{65}$$

the former by construction (see (26)).

The variation of $\Lambda^k(u^k, p^k)$ over one stage of the algorithm is

$$
(66)\qquad \left\{\begin{aligned}
\delta_k^{k+1}\Lambda = {} & \underbrace{K(u^k)-K(u^{k+1})-\langle K'(u^k), u^k-u^{k+1}\rangle}_{A_1} \\
& +\underbrace{\langle K'(u^k)-K'(u^{k+1}), u^*-u^{k+1}\rangle}_{A_2} \\
& +\underbrace{(\varepsilon^{k+1}\|p^{k+1}-p^*\|^2-\varepsilon^k\|p^k-p^*\|^2)/2\rho.}_{A_3}
\end{aligned}\right.
$$

The inequalities (57) for A_1 and (58) for A_2 are still valid (remember that $g\equiv 0$, $\Xi\equiv 0$). As for A_3, we get

$$
\begin{aligned}
A_3 &\leqq (\varepsilon^k/2\rho)(\|p^{k+1}-p^*\|^2-\|p^k-p^*\|^2) \\
&\leqq \varepsilon^k\langle p^k-p^*, \Theta(u^{k+1})-\Theta(u^*)\rangle+\varepsilon^k\rho\tau^2\|u^{k+1}-u^*\|^2/2,
\end{aligned}
$$

the former because $\{\varepsilon^k\}$ is nonincreasing, and the latter thanks to (26) and (60), but with Π_μ instead of Π—this holds true because we assume that $p^*\in B(0,\mu)$—and thanks to the Lipschitz property of Θ.

Summing up these inequalities for A_1, A_2, A_3 and using the convexity of Θ and the definition of θ^k, the Schwarz inequality and the Lipschitz property of Θ (also $\theta^k\leqq\tau$), we get

$$
\begin{aligned}
\delta_k^{k+1}\Lambda \leqq {} & \varepsilon^k[\langle r^k, u^*-u^k\rangle+\langle p^*, \Theta(u^*)-\Theta(u^k)\rangle] \\
& +\underbrace{\varepsilon^k[\|r^k\|+2\tau\|p^k\|+\tau p^*]\cdot\|u^{k+1}-u^k\|}_{B_1} \\
& +\underbrace{\varepsilon^k\rho\tau^2\|u^{k+1}-u^*\|^2/2-b\|u^{k+1}-u^k\|^2/2.}_{B_2}
\end{aligned}
$$

The last terms above can be bounded as follows. Remembering (12), (55), the fact that $\|p^k\|\leqq\mu$ and using standard manipulations, we get

$$
(67)\qquad \begin{aligned}
B_1 &\leqq (\varepsilon^k)^2[c_5\|u^k-u^*\|^2+c_6] \\
&\leqq (\varepsilon^k)^2[c_7\Lambda^k(u^k,p^k)+c_6],
\end{aligned}
$$

the latter from (65). On the other hand, note that

$$
\begin{aligned}
\|u^{k+1}-u^*\|^2 &\leqq \|u^k-u^*\|^2+\|u^{k+1}-u^k\|^2+2\|u^k-u^*\|\cdot\|u^{k+1}-u^k\| \\
&\leqq (1+\alpha\varepsilon^k)\|u^k-u^*\|^2+(1+1/\alpha\varepsilon^k)\|u^{k+1}-u^k\|^2
\end{aligned}
$$

from the Hölder inequality

$$
(68)\qquad xy\leqq(\xi x^2+y^2/\xi)/2,
$$

which is true for an arbitrary positive number ξ. Hence

$$
\begin{aligned}
B_2 \leqq {} & \underbrace{\varepsilon^k\rho\tau^2\|u^k-u^*\|^2/2}_{C_1}+\underbrace{(\varepsilon^k)^2\alpha\rho\tau^2\|u^k-u^*\|^2/2}_{C_2} \\
& +\underbrace{(\rho\tau^2(\varepsilon^k+1/\alpha)-b)\|u^{k+1}-u^k\|^2/2.}_{C_3}
\end{aligned}
$$

If we choose α as the solution of $\rho\tau^2(\varepsilon^0+1/\alpha)=b$, then clearly C_3 will always be nonnegative, whereas $C_2 \leqq (\varepsilon^k)^2[c_8\Lambda^k(u^k, p^k)+c_9]$ from (65). This last bound is similar to (67).

Collecting everything, we get

$$\delta_k^{k+1}\Lambda \leqq \varepsilon^k[\langle r^k, u^*-u^k\rangle+\langle p^*, \Theta(u^*)-\Theta(u^k)\rangle] + (\varepsilon^k)^2[c_8\Lambda^k(u^k, p^k)+c_9]+\varepsilon^k\rho\tau^2\|u^k-u^*\|^2/2.$$

Taking the conditional expectation knowing $\mathfrak{F}^k$—remembering (13)—yields

$$\mathbb{E}^k\Lambda^{k+1}(u^{k+1}, p^{k+1})-\Lambda^k(u^k, p^k) \leqq \varepsilon^k[\langle \mathbb{E}^k r^k, u^*-u^k\rangle+\langle p^*, \Theta(u^*)-\Theta(u^k)\rangle] + (\varepsilon^k)^2[c_8\Lambda^k(u^k, p^k)+c_9]+\varepsilon^k\rho\tau^2\|u^k-u^*\|^2/2.$$

Observe that $R^k := \mathbb{E}^k r^k = \mathbb{E} r^k \in \partial J(u^k)$ (similar to (14)). Moreover

$$\langle R^k, u^*-u^k\rangle \leqq \langle R^*, u^*-u^k\rangle - a\|u^k-u^*\|^2$$

(see footnote 6) for all $R^* \in \partial J(u^*)$. Finally, because u^* minimizes $J(\cdot)+\langle p^*, \Theta(\cdot)\rangle$, there exists some $R^* \in \partial J(u^*)$ such that

$$\forall u \in U^f, \quad \langle R^*, u-u^*\rangle+\langle p^*, \Theta(u)-\Theta(u^*)\rangle \geqq 0.$$

These considerations yield the basic inequality

$$\mathbb{E}^k\Lambda^{k+1}(u^{k+1}, p^{k+1})-\Lambda^k(u^k, p^k) \leqq \varepsilon^k\left(\frac{\rho\tau^2}{2}-a\right)\|u^k-u^*\|^2+(\varepsilon^k)^2[c_8\Lambda^k(u^k, p^k)+c_9]$$

which plays the part of (50) in the proof of Theorem 1. Remembering (27), it can be proved in the same way as previously that $\|u^k-u^*\| \to 0$ as $k \to +\infty$.

D. Proof of Theorem 4. For the sake of brevity, let us set

$$q^k := \Pi[p^k + c\Theta(u^k)] \qquad (=(\lambda_c)_t'(\Theta(u^k), p^k)) \tag{69}$$

$$q^{k+1/2} := \Pi[p^k + c\Theta(u^{k+1})] \qquad (=(\lambda_c)_t'(\Theta(u^{k+1}), p^k)). \tag{70}$$

The variational inequality (54) now holds with $g \equiv 0$, $\Xi \equiv 0$, and with q^k replacing p^k.

We use the Lyapounov function (64) again and thus (65) is still valid. However, the proof will be a bit more involved in that it will require two stages. These stages correspond to Lemmas 7 and 8 hereafter. But we start with Lemma 6 which is a technical lemma about function λ_c.

LEMMA 6. *For function λ_c defined by* (21), *we have*

$$\overbrace{\langle (\lambda_c)_t'(t_1, p_1)-(\lambda_c)_t'(t_2, p_2), t_1-t_2\rangle - \langle (\lambda_c)_p'(t_1, p_1)-(\lambda_c)_p'(t_2, p_2), p_1-p_2\rangle}^{Y} \geqq c\|(\lambda_c)_p'(t_1, p_1)-(\lambda_c)_p'(t_2, p_2)\|^2. \tag{71}$$

Proof. From (22)-(23), we have that

$$\begin{aligned} Y &= \langle \Pi(p_1+ct_1)-\Pi(p_2+ct_2), c(t_1-t_2)\rangle/c \\ &\quad -\langle \Pi(p_1+ct_1)-p_1-(\Pi(p_2+ct_2)-p_2), p_1-p_2\rangle/c \\ &= \|p_1-p_2\|^2/c - 2\langle \Pi(p_1+ct_1)-\Pi(p_2+ct_2), p_1-p_2\rangle/c \\ &\quad +\langle \Pi(p_1+ct_1)-\Pi(p_2+ct_2), p_1+ct_1-(p_2+ct_2)\rangle/c. \end{aligned}$$

The last term is not less than $\|\Pi(p_1+ct_1)-\Pi(p_2+ct_2)\|^2/c$ because in general

$$\|\Pi a-\Pi b\|^2 \leqq \langle \Pi a-\Pi b, a-b\rangle.$$

Therefore

$$Y \geqq \|\Pi(p_1+ct_1)-p_1-(\Pi(p_2+ct_2)-p_2)\|^2/c,$$

which is the claimed result. □

LEMMA 7. *Almost surely*

$$\sum_{k\in\mathbb{N}} \varepsilon^k \mathbb{E}\|q^{k+1/2}-p^k\|^2 < \infty. \tag{72}$$

Proof. We decompose the variation of the Lyapounov function in three terms as in (66). Inequality (57) still holds true whereas (58) now reads

$$\begin{aligned} A_2 &\leqq \varepsilon^k[\langle r^k, u^*-u^{k+1}\rangle + \langle q^k, \theta^k\cdot(u^*-u^{k+1})\rangle] \\ &\leqq \varepsilon^k[j(u^*,\omega^{k+1})-j(u^k,\omega^{k+1}) && \text{(convexity of } j(\cdot,\omega^{k+1}))\\ &\quad + \langle q^k, \Theta(u^*)-\Theta(u^{k+1})\rangle && \text{(convexity of } \Theta)\\ &\quad + \|r^k\|\cdot\|u^{k+1}-u^k\|] && \text{(Schwarz inequality).} \end{aligned}$$

As for A_3, we have that

$$\begin{aligned} A_3 &\leqq (\varepsilon^k/2\rho)(\|p^{k+1}-p^*\|^2-\|p^k-p^*\|^2) && (\{\varepsilon^k\} \text{ nonincreasing})\\ &\leqq \varepsilon^k\langle p^k-p^*, (q^{k+1/2}-p^k)/c\rangle \\ &\quad + (\varepsilon^k\rho/2c^2)\|q^{k+1/2}-p^k\|^2 && \text{(from (31) and } p^*\in B(0,\mu)). \end{aligned}$$

Summing up

$$\begin{aligned} \delta_k^{k+1}\Lambda &\leqq \varepsilon^k[\ell(u^*,p^*,\omega^{k+1})-\ell(u^k,p^*,\omega^{k+1})] && \text{(see (61))}\\ &\quad \underbrace{+\varepsilon^k[\langle q^{k+1/2}-p^*, \Theta(u^*)-\Theta(u^{k+1})\rangle + \langle (q^{k+1/2}-p^k)/c, p^k-p^*\rangle]}_{B_1}\\ &\quad \underbrace{+\varepsilon^k[\langle q^k-q^{k+1/2}, \Theta(u^*)-\Theta(u^{k+1})\rangle + \langle p^*, \Theta(u^k)-\Theta(u^{k+1})\rangle]}_{B_2}\\ &\quad \underbrace{+\varepsilon^k\|r^k\|\cdot\|u^{k+1}-u^k\|}_{B_3} + (\varepsilon^k\rho/2c^2)\|q^{k+1/2}-p^k\|^2 - b\|u^{k+1}-u^k\|^2/2. \end{aligned}$$

For B_1, we use Lemma 6 with $t_1=\Theta(u^{k+1})$, $t_2=\Theta(u^*)$, $p_1=p^k$, $p_2=p^*$. With these choices, remembering (22)-(23) and noting that $(\lambda_c)_t'(t_2,p_2)=p^*$ and that $(\lambda_c)_p'(t_2,p_2)=0$ since $p^*=\Pi(p^*+c\Theta(u^*))$, it is realized that B_1 is exactly the expression $-Y$ in that lemma. Hence

$$B_1 \leqq -\|q^{k+1/2}-p^k\|^2/c.$$

For B_2, we simply use (69)-(70) and the Lipschitz property of Θ,

$$\begin{aligned} B_2 &\leqq \tau^2\|u^{k+1}-u^k\|\cdot\|u^{k+1}-u^*\| + \tau\|p^*\|\cdot\|u^{k+1}-u^k\| \\ &\leqq \|u^{k+1}-u^k\|\cdot[\tau^2(\|u^{k+1}-u^k\|+\|u^k-u^*\|)+\tau\|p^*\|] && \text{(triangular inequality)}\\ &\leqq \left(\tau^2+\frac{\tau^2+\tau}{2\alpha\varepsilon^k}\right)\|u^{k+1}-u^k\|^2 + \frac{\tau^2\alpha\varepsilon^k}{2}\|u^k-u^*\|^2 + \frac{\tau\alpha\varepsilon^k}{2}\|p^*\|^2 \end{aligned}$$

for any positive number α using (68). For the same reason,

$$\begin{aligned} B_3 &\leqq \frac{1}{2\alpha\varepsilon^k}\|u^{k+1}-u^k\|^2 + \frac{\alpha\varepsilon^k}{2}\|r^k\|^2 \\ &\leqq \frac{1}{2\alpha\varepsilon^k}\|u^{k+1}-u^k\|^2 + \varepsilon^k(c_5\|u^k-u^*\|^2+c_6) \end{aligned}$$

using (12).

Collecting these inequalities, we get

$$\delta_k^{k+1}\Lambda \leqq \varepsilon^k[\ell(u^*, p^*, \omega^{k+1}) - \ell(u^k, p^*\omega^{k+1})]$$
$$+\frac{\varepsilon^k}{c}\left(\frac{\rho}{2c}-1\right)\|q^{k+1/2}-p^k\|^2+(\varepsilon^k)^2(c_7\|u^k-u^*\|^2+c_8)$$
$$+\left(\frac{c_9}{\alpha}+c_{10}\varepsilon^k-\frac{b}{2}\right)\|u^{k+1}-u^k\|^2$$

where c_7 and c_8 depend on α. This α is chosen in such a way that

$$c_9/\alpha + c_{10}\varepsilon^0 - b/2 = 0$$

so that the last term is nonpositive for all k (remember that $\{\varepsilon^k\}$ is a nonincreasing sequence).

Then taking the conditional expectation knowing $\mathfrak{F}^k$, we get

$$\mathbb{E}^k\Lambda^{k+1}(u^{k+1}, p^{k+1}) - \Lambda^k(u^k, p^k) \leqq \varepsilon^k[L(u^*, p^*) - L(u^k, p^*)]$$
$$+\frac{\varepsilon^k}{c}\left(\frac{\rho}{2c}-1\right)\mathbb{E}^k\|q^{k+1/2}-p^k\|^2$$
$$+(\varepsilon^k)^2\left(\frac{2c_7}{b}\Lambda^k(u^k, p^k)+c_8\right)$$

where we have used (65). Observe that the first term in the right-hand side is nonnegative. This is also the case for the second term from (32). We can thus drop them and take the expectation to get (51) again (here with $y^k := \mathbb{E}\Lambda^k(u^k, p^k)$), from which we conclude that $\{\Lambda^k(u^k, p^k)\}$ is a quasimartingale by the same argument as in the proof of Theorem 1. Then (72) follows similarly to (52). □

LEMMA 8. *Almost surely*

$$\sum_{k\in\mathbb{N}} \varepsilon^k(L_c(u^k, p^*) - L_c(u^*, p^k)) < \infty. \tag{73}$$

Proof. We return to the inequalities obtained for A_2 and A_3 at the beginning of the proof of Lemma 7. Because of (69) and the convexity of $\lambda_c(\cdot, p^k)$, we have that

$$A_2 \leqq \varepsilon^k[j(u^*, \omega^{k+1}) - j(u^k, \omega^{k+1}) + \lambda_c(\Theta(u^*), p^k) - \lambda_c(\Theta(u^k), p^k)$$
$$+(\|r^k\|+\tau\|q^k\|)\cdot\|u^{k+1}-u^k\|].$$

Because $(q^k - p^k)/c = (\lambda_c)'_p(\Theta(u^k), p^k)$ and because of the concavity of $\lambda_c(\Theta(u^k), \cdot)$, we have that

$$A_3 \leqq \varepsilon^k[\lambda_c(\Theta(u^k), p^k) - \lambda_c(\Theta(u^k), p^*) + (\rho/2c^2)\|q^{k+1/2}-p^k\|^2$$
$$+\|p^k-p^*\|\cdot\|q^{k+1/2}-q^k\|].$$

If we collect these new inequalities for A_2 and A_3 together with (57) for A_1, we obtain

$$\delta_k^{k+1}\Lambda \leqq \varepsilon^k[\ell_c(u^*, p^k, \omega^{k+1}) - \ell_c(u^k, p^*, \omega^{k+1}) + (\rho/2c^2)\|q^{k+1/2}-p^k\|^2]$$
$$\underbrace{+\varepsilon^k[(\|r^k\|+\tau\|q^k\|)\|u^{k+1}-u^k\| + \|p^k-p^*\|\cdot\|q^{k+1/2}-q^k\|] - b\|u^{k+1}-u^k\|^2/2}_{B_4}$$

where we have set, similarly to (61)

$$\ell_c(u, p, \omega) := j(u, \omega) + \lambda_c(\Theta(u), p).$$

From (69)-(70) and the Lipschitz property of Θ, it comes that

$$\|q^{k+1/2}-q^k\| \leqq \tau\|u^{k+1}-u^k\|.$$

Hence using (68) again, for any positive number β

$$B_4 \leqq \frac{\beta(\varepsilon^k)^2}{2}[\|r^k\|^2+\tau(\|q^k\|^2+\|p^k-p^*\|^2)]$$
$$+\left(\frac{1+2\tau}{2\beta}-\frac{b}{2}\right)\|u^{k+1}-u^k\|^2.$$

From (69), we have that

$$\begin{aligned}\|q^k\| &\leqq \|p^k\|+c\|\Theta(u^*)\|+c\|\Theta(u^k)-\Theta(u^*)\| \\ &\leqq \|p^k\|+c\|\Theta(u^*)\|+c\tau\|u^k-u^*\|.\end{aligned} \tag{74}$$

We use this inequality for q^k, (12) for r^k and (65) for p^k and u^k-u^*, plus standard calculations. On the other hand, we choose β in such a way that $(1+2\tau)/2\beta = b/2$. Thus we obtain

$$B_4 \leqq (\varepsilon^k)^2[c_{11}\Lambda^k(u^k,p^k)+c_{12}].$$

A new inequality follows for $\delta_k^{k+1}\Lambda$ on which we take the conditional expectation knowing $\mathfrak{F}^k$—in particular $\mathbb{E}^k[\ell_c(u^*,p^k,\omega^{k+1})-\ell_c(u^k,p^*,\omega^{k+1})] = L_c(u^*,p^k)-L_c(u^k,p^*)$—and then the expectation. This yields

$$\begin{aligned}\varepsilon^k[L_c(u^k,p^*)-L_c(u^*,p^*)] \leqq y^k-y^{k+1}+(\varepsilon^k)^2[c_{11}y^k+c_{12}] \\ +(\varepsilon^k\rho/2c^2)\mathbb{E}\|q^{k+1/2}-p^k\|^2.\end{aligned}$$

Recall that $y^k := \mathbb{E}\Lambda^k(u^k,p^k)$ is almost surely bounded. Let Y be a bound, then summing up for $k=0,\cdots,N-1$, we get

$$\sum_{k=0}^{N-1}\varepsilon^k[L_c(u^k,p^*)-L_c(u^*,p^*)] \leqq \overbrace{y_0-y^N+\sum_{k=0}^{N-1}(\varepsilon^k)^2[c_{11}y^k+c_{12}]}^{B_5}$$
$$\underbrace{+\sum_{k=0}^{N-1}(\varepsilon^k\rho/2c^2)\mathbb{E}\|q^{k+1/2}-p^k\|^2}_{B_6}$$

and

$$B_5 \leqq Y+\sum_{k=0}^{N-1}(\varepsilon^k)^2[c_{11}Y+c_{12}] \leqq Y+\sum_{k\in\mathbb{N}}(\varepsilon^k)^2[c_{11}Y+c_{12}]<\infty$$

whereas B_6 is also bounded from (72). We conclude that (73) holds true almost surely since the expectation can be removed for $L_c(u^k,p^*)-L_c(u^*,p^*)\geqq 0$ for all k. □

We can now complete the proof of Theorem 4. In the same way that we derived (55), we can now obtain

$$\begin{aligned}\|u^{k+1}-u^k\| &\leqq \varepsilon^k[c_{13}\|u^k-u^*\|+c_{14}\|q^k\|+c_{15}] \\ &\leqq \varepsilon^k[c_{16}\|u^k-u^*\|+c_{17}] && \text{(from (74))} \\ &\leqq \varepsilon^k c_{18} && \text{(a.s.)}\end{aligned}$$

since we have seen that $\{u^k\}$ is almost surely bounded in the proof of Lemma 7. A similar statement cannot be obtained for $p^{k+1}-p^k$ due to the use of "large steps" to update p. For this reason, we cannot exploit (73) and we limit ourselves to the weaker result

$$\sum_{k\in\mathbb{N}} \varepsilon^k(L_c(u^k, p^*)-L_c(u^*, p^*))<\infty \quad \text{a.s.}$$

which follows from (73) since $L_c(u^*, p^*)\geqq L_c(u^*, p^k)$. Given that $L_c(\cdot, p^*)$ is Lipschitz on every bounded set, we conclude from Lemma 4 of [13] that

$$\lim_{k\to+\infty} L_c(u^k, p^*)=L_c(u^*, p^*) \quad \text{a.s.}$$

Since the sequence $\{u^k\}$ is almost surely bounded, it has cluster points in the weak topology, and for each such cluster point $\bar{u}$, $L_c(\bar{u}, p^*)=\min_{u\in U^f} L_c(u, p^*)$ almost surely since $L_c(\cdot, p^*)$ is weakly lower semicontinuous. This implies that every $\bar{u}$ is almost surely a solution of (9). This holds true for the augmented Lagrangian (this property is sometimes referred to as "stability in u"—see [13]) but not for the ordinary Lagrangian in general, unless it is *strictly* convex (the case considered in Theorem 2).

Finally, if J is strongly convex, it can be proved as usual that the convergence of u^k towards the unique u^* is strong.

E. Proof of Theorem 5. The claims at (i) and (ii) should be clear given assumptions. A primal-dual solution of (40) is characterized by the conditions

$$\forall u\in\mathcal{U}, \quad \langle K'(u^{k+1})-K'(u^k)+\varepsilon^k(r^k+E^Tp^{k+1}), u-u^{k+1}\rangle + \varepsilon^k(g(u,\omega)^{k+1})-g(u^{k+1},\omega^{k+1}))\geqq 0 \tag{75}$$

$$Eu^{k+1}=v^k. \tag{76}$$

Let us use (75) with $u=u^{k+1}-xE^Tp^{k+1}$ where x is an arbitrary positive constant to be chosen later on. Let e be the operator norm of E^T or E (which is bounded since it is linear and continuous). We get that

$$\begin{aligned} x\varepsilon^k\|E^Tp^{k+1}\|^2 &\leqq \langle K'(u^{k+1})-K'(u^k)+\varepsilon^k r^k, -x\varepsilon^k E^Tp^{k+1}\rangle \\ &\quad +\varepsilon^k(g(u,\omega^{k+1})-g(u^{k+1},\omega^{k+1})) \\ &\leqq x\,e\|p^{k+1}\|[B\|u^{k+1}-u^k\|+\varepsilon^k\|r^k\| \quad \text{(Schwarz and Lipschitz for } K') \\ &\quad +\varepsilon^k(c_5(\|u^{k+1}\|+x\,e\|p^{k+1}\|)+c_6)] \qquad \text{(from property (16) for } g). \end{aligned}$$

Since E is assumed to be onto, we know that (see Remark 11)

$$\exists\delta>0: \forall p\in\mathcal{C}^*, \quad \|E^Tp\|^2=\langle p, EE^Tp\rangle\geqq\delta\|p\|^2.$$

Therefore, from the previous inequality, we derive the following:

$$\delta\|p^{k+1}\|\leqq e[B\|u^{k+1}-u^k\|/\varepsilon^k+\|r^k\|+c_5(\|u^{k+1}\|+x\,e\|p^{k+1}\|)+c_6].$$

Since this is true for an arbitrarily small x, the same inequality holds true without the term $x\,e\|p^{k+1}\|$ in the right-hand side. Moreover, using (12) and $\|u^{k+1}\|\leqq\|u^k\|+\|u^{k+1}-u^k\|$, we get

$$\|p^{k+1}\|\leqq(e/\delta)(c_5+B/\varepsilon^k)\|u^{k+1}-u^k\|+c_7\|u^k\|+c_8. \tag{77}$$

Let us introduce the Lyapounov function

$$\Lambda(u, v)=K(u^*)-K(u)-\langle K'(u), u^*-u\rangle+\|v-v^*\|^2/2\gamma$$

where u^* is a solution of (38) and

$$v^*=Eu^*. \tag{78}$$

Because K is strongly convex,

$$\|u^k-u^*\|^2 \leqq 2\Lambda(u^k, v^k)/b \quad \text{and} \quad \|v^k-v^*\|^2 \leqq 2\gamma\Lambda(u^k, v^k). \tag{79}$$

We consider the variation of Λ over one iteration of the algorithm

$$\delta_k^{k+1}\Lambda = \underbrace{K(u^k)-K(u^{k+1})-\langle K'(u^k), u^k-u^{k+1}\rangle}_{A_1}$$
$$+\underbrace{\langle K'(u^k)-K'(u^{k+1}), u^*-u^{k+1}\rangle}_{A_2}$$
$$+\underbrace{(\|v^{k+1}-v^*\|^2-\|v^k-v^*\|^2)/2\gamma.}_{A_3}$$

For A_1, inequality (57) holds true. For A_2, using (75) with $u=u^*$, we get something similar to (59), which reads, remembering (76) and (78)

$$A_2 \leqq \varepsilon^k[(j+g)(u^*, \omega^{k+1})-(j+g)(u^k, \omega^{k+1})+\langle p^{k+1}, v^*-v^k\rangle]$$
$$+\underbrace{\varepsilon^k[\|r^k\|+c_5(\|u^k\|+\|u^{k+1}-u^k\|)+c_6]\cdot\|u^{k+1}-u^k\|}_{B_1}.$$

But

$$B_1 \leqq \varepsilon^k c_5\|u^{k+1}-u^k\|^2+\varepsilon^k\|u^{k+1}-u^k\|\cdot(c_{11}\|u^k\|+c_{12}) \qquad \text{(from (12))}$$
$$\leqq (\varepsilon^k c_5+\alpha/2)\|u^{k+1}-u^k\|^2+(\varepsilon^k)^2(c_{11}\|u^k\|+c_{12})^2/2\alpha \qquad \text{((68) with any } \alpha>0)$$
$$\leqq (\varepsilon^k c_5+\alpha/2)\|u^{k+1}-u^k\|^2+(\varepsilon^k)^2(c_{13}\Lambda(u^k, v^k)+c_{14})$$

using $(x+y)^2 \leqq 2(x^2+y^2)$ and (79).

As for A_3, with help of (41) and the fact that $v^* \in V^f$, we have that

$$A_3 \leqq \varepsilon^k\langle p^{k+1}, v^k-v^*\rangle+\underbrace{\gamma(\varepsilon^k)^2\|p^{k+1}\|^2/2.}_{B_2}$$

However, from (77) and the fact that $(x+y)^2 \leqq (1+\alpha)x^2+(1+1/\alpha)y^2$ for any $\alpha>0$,

$$B_2 \leqq \gamma(1+\alpha)\, e^2(B+c_5\varepsilon^k)^2\|u^{k+1}-u^k\|^2/2\delta^2+\gamma(1+1/\alpha)(\varepsilon^k)^2(c_7\|u^k\|+c_8)^2/2$$
$$\leqq \gamma(1+\alpha)\, e^2(B+c_5\varepsilon^k)^2\|u^{k+1}-u^k\|^2/2\delta^2+(\varepsilon^k)^2(c_{15}\Lambda(u^k, v^k)+c_{16})$$

using (79) again.

Collecting everything, we get

$$\delta_k^{k+1}\Lambda \leqq \varepsilon^k[(j+g)(u^*, \omega^{k+1})-(j+g)(u^k, \omega^{k+1})]+(\varepsilon^k)^2[c_{17}\Lambda(u^k, v^k)+c_{18}]$$
$$+[\varepsilon^k c_5+\alpha/2+\gamma(1+\alpha)\, e^2(B+c_5\varepsilon^k)^2/2\delta^2-b/2]\cdot\|u^{k+1}-u^k\|^2.$$

Clearly, with α and γ small enough and when ε^k has reached a sufficiently small value also, (say, it is reached already for $k=0$), the coefficient in front of $\|u^{k+1}-u^k\|^2$ becomes nonpositive.

Remark 13. Indeed, α can be chosen arbitrarily small so that γ has to be less than a certain fraction of the ratio $z=b\delta^2/e^2B^2$. Note that if E is changed into νE, then δ is changed into $\nu^2\delta$ whereas e is changed into νe. In the same way, if K is divided by ν', both b and B are changed in the same way, so that z is finally multiplied by $\nu^2\nu'$. In this way, the bound on γ can be made arbitrarily large.

Then, taking the conditional expectation

$$\mathbb{E}^k\Lambda(u^{k+1}, v^{k+1})-\Lambda(u^k, v^k) \leqq \varepsilon^k[(j+G)(u^*)-(J+G)(u^k)]+(\varepsilon^k)^2[c_{17}\Lambda(u^k, v^k)+c_{18}],$$

observe that the first term in the right-hand side is nonpositive since u^k is feasible for (38). Using the arguments already advocated in the previous proofs, it can then be proved that

- $\{\Lambda(u^k, v^k)\}$ is a quasimartingale which converges almost surely to some variable with finite expectation:
- the sequences $\{u^k\}$ and $\{v^k\}$ are almost surely bounded from (79);
- one has that

$$\sum_{k\in\mathbb{N}} \varepsilon^k[(J+G)(u^k)-(J+G)(u^*)]<+\infty \quad \text{a.s.} \tag{80}$$

Suppose that we show that

$$\exists c_{19}: \forall k\in\mathbb{N}, \quad \|u^{k+1}-u^k\|\leqq c_{19}\varepsilon^k \quad \text{a.s.} \tag{81}$$

which we postpone to the end of this proof. Then, applying Lemma 4 of [13] once again, since $J+G$ is Lipschitz on the almost surely bounded set containing the sequence $\{u^k\}$, with (80) we conclude that $(J+G)(u^k)\to(J+G)(u^*)$ almost surely. Therefore, using the (weak) lower semicontinuity of $J+G$, each cluster point $\bar{u}$ of $\{u^k\}$ is such that $(J+G)(\bar{u})\leqq(J+G)(u^*)$. Clearly, the set of feasible points of (38) is closed and convex hence weakly closed. Thus, $\bar{u}$ is also feasible, hence optimal.

The case when $J+G$ is *strongly* convex is handled classically.

We now complete the proof by proving (81). Note that this will imply the boundedness of $\{p^k\}$ almost surely from (77). We appeal to the variational inequality (75) once again, but now with $u=u^k$. This yields

$$\begin{aligned} b\|u^{k+1}-u^k\|^2 &\leqq \langle K'(u^{k+1})-K'(u^k), u^{k+1}-u^k\rangle \\ &\leqq \varepsilon^k[\langle r^k, u^k-u^{k+1}\rangle + g(u^k,\omega^{k+1})-g(u^{k+1},\omega^{k+1})] \\ &\quad + \varepsilon^k\langle p^{k+1}, Eu^k-Eu^{k+1}\rangle \\ &\leqq \varepsilon^k[\|r^k\|+c_5(\|u^k\|+\|u^{k+1}-u^k\|)+c_6]\cdot\|u^{k+1}-u^k\| \\ &\quad + \varepsilon^k\|p^{k+1}\|\cdot\|v^{k-1}-v^k\| \end{aligned}$$

using the Schwarz inequality and (16) for g, and (76) for v^k and v^{k-1}. On the one hand, u^k, hence also r^k, are bounded almost surely, on the other hand, from (41)

$$\|v^k-v^{k-1}\|\leqq\gamma\varepsilon^{k-1}\|p^k\|\leqq\gamma\zeta\varepsilon^k\|p^k\|$$

using (42). Therefore

$$\left(\frac{\|u^{k+1}-u^k\|}{\varepsilon^k}\right)^2\leqq\frac{1}{b}\left(c_5\varepsilon^k\frac{\|u^{k+1}-u^k\|}{\varepsilon^k}+c_{20}\right)\frac{\|u^{k+1}-u^k\|}{\varepsilon^k}+\frac{\gamma\zeta}{b}\|p^k\|\cdot\|p^{k+1}\|.$$

Knowing that u^k is almost surely bounded, (77) yields

$$\|p^{k+1}\|\leqq\frac{e}{\delta}(c_5\varepsilon^k+B)\frac{\|u^{k+1}-u^k\|}{\varepsilon^k}+c_{21}.$$

To shorten notations, we set

$$x^{k+1}=\|u^{k+1}-u^k\|/\varepsilon^k \quad \text{and} \quad y^{k+1}=\|p^{k+1}\|.$$

The two inequalities above then imply

$$\begin{aligned} (x^{k+1})^2 &\leqq c_{22}(x^{k+1})^2+c_{23}x^{k+1}+\gamma c_{24}y^1y^{k+1} \\ y^{k+1} &\leqq c_{25}x^{k+1}+c_{21} \end{aligned}$$

where $c_{23} = c_{20}/b$, $c_{24} = \zeta/b$ and c_{22} (respectively, c_{25}) is larger but as close as we wish to zero (respectively, eB/δ) provided that we consider these inequalities from some k large enough. Using the latter inequality in the former yields

$$(x^{k+1})^2 \leqq c_{22}(x^{k+1})^2 + c_{23}x^{k+1} + \gamma c_{24}(c_{25}x^{k+1} + c_{21})(c_{25}x^k + c_{21}).$$

After tedious but straightforward calculations where we use (68) repeatedly (always with $\xi = \alpha$ except for the cross-product $x^k x^{k+1}$ for which we take $\xi = \beta$), we finally get

$$(1 - \beta\gamma c_{24}c_{25}^2/2 - c_{22} - \alpha c_{23}/2 - \alpha\gamma c_{21}c_{24}c_{25}/2)(x^{k+1})^2$$
$$\leqq (\gamma c_{24}c_{25}^2/2\beta + \alpha\gamma c_{21}c_{24}c_{25}/2)(x^k)^2 + c_{23}/2\alpha + \gamma c_{21}c_{24}c_{25}/\alpha + \gamma c_{24}c_{21}^2.$$

To fix ideas, let us pick $\beta = 1/\gamma c_{24}c_{25}^2$. It should be clear that since c_{22} is small for k large and since α can be chosen small also, this inequality can assume the following form:

$$(x^{k+1})^2 \leqq c_{26}(x^k)^2 + c_{27}$$

with $c_{26} < 1$. By a classical result, this proves that the sequence $\{x^k\}$ is bounded, which is the desired result.

A careful examination shows once again that if α is chosen arbitrarily small, γ must be chosen less than a certain fraction of the ratio $z = b\delta^2/e^2B^2$ as already noticed in Remark 13, which thus remains valid.

REFERENCES

[1] K. J. ARROW, L. HURWICZ, AND H. UZAWA, *Studies in Linear and Non-Linear Programming*, Stanford University Press, CA, 1972.

[2] J. F. BALDUCCHI, G. COHEN, J. C. DODU, M. GOURSAT, M. HERTZ, J. P. QUADRAT, AND M. VIOT, *Three methods for optimizing the capacities of an electrical transmission network*, IFAC World Congress, Kyoto, Japan, 1981.

[3] A BENVENISTE, P. BERNHARD, AND G. COHEN, *On the decomposition of stochastic control problems*, 1st IFAC Symp. on Large Scale Systems Theory and Applications, Udine, Italy, 1976.

[4] D. P. BERTSEKAS, *Necessary and sufficient conditions for a penalty method to be exact*, Math. Programming, 9 (1975), pp. 87–99.

[5] C. B. BROSILOW, L. S. LASDON, AND J. D. PEARSON, *Feasible optimization methods for interconnected systems*, Joint Automatic Control Conference, Troy, New York, 1965.

[6] G. COHEN, *Optimization by decomposition and coordination: a unified approach*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 222–232.

[7] ———, *Auxiliary problem principle and decomposition of optimization problems*, J. Optim. Theory Appl., 32 (1980), pp. 277–305.

[8] ———, *Two lemmas and their use in convergence analysis of some optimization algorithms*, Int. rep. E/68, Centre d'Automatique et Informatique, École des Mines de Paris, Fontainebleau, France, 1982.

[9] ———, *Décomposition et coordination en optimisation déterministe différentiable et non différentiable*, Thesis dissertation, University of Paris-Dauphine, Paris, France, 1984.

[10] ———, *Nash equilibria: gradient and decomposition algorithms*, Large Scale Systems, 12 (1987), pp. 173–184.

[11] G. COHEN, *Auxiliary problem principle extended to variational inequalities*, J. Optim. Theory Appl., 59 (1988), pp. 325–333.

[12] G. COHEN AND J. C. CULIOLI, *Decomposition and coordination in stochastic optimization*, IFAC Symp. on Large Scale Systems Theory and Applications, Zurich, Switzerland, 1986.

[13] G. COHEN AND D. L. ZHU, *Decomposition coordination methods in large scale optimization problems. The nondifferentiable case and the use of augmented Lagrangians*, in Advances in Large Scale Systems Theory and Applications, Vol. I, J. B. Cruz, ed., JAI Press, Greenwich, Connecticut, 1984.

[14] J. C. CULIOLI *Algorithmes de décomposition/coordination en optimisation stochastique*, Thesis dissertation, Centre d'Automatique et Informatique, École des Mines de Paris, Fontainebleau, France, 1987.

[15] G. B. DANTZIG AND P. WOLFE, *The decomposition algorithm for linear program*, Econometrica, 29 (1961), pp. 767–778.

[16] Y. M. ERMOLIEV, *Methods of Stochastic Programming*, Nauka, Moscow, 1976. (In Russian.)

[17] ———, *Stochastic quasigradient methods and their applications in systems optimization*, IIASA working paper WP-81-2, Laxenburg, Austria, January 1981.

[18] A. M. GEOFFRION, *Primal resource-directive approaches for optimizing nonlinear decomposable systems*, Oper. Res. (1970), pp. 375–403.

[19] A. M. GUPAL, *Stochastic Solution Methods for Nonsmooth Minimax Problems*, Naukova Dunka, Kiev, 1979. [In Russian.]

[20] H. KUSHNER AND D. S. CLARK, *Stochastic Approximation for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, 1978.

[21] L. S. LASDON AND J. D. SCHOEFFLER, *A multilevel technique for optimization*, Joint Automatic Control Conference, Troy, New York, 1965.

[22] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551–575.

[23] D. P. LOOZE AND N. R. SANDELL JR., *Decomposition of linear decentralized stochastic control problems*, IFAC Workshop on Control and Management of Integrated Industrial Complexes, Toulouse, France, 1977.

[24] F. V. LOUVEAUX, *A solution method for multistage stochastic programming with recourse, with application to an energy investment problem*, Oper. Res., 28 (1980), pp. 889–902.

[25] M. D. MESAROVIC, D. MACKO, AND Y. TAKAHARA, *Theory of Hierarchical Multilevel Systems*, Academic Press, New York, 1970.

[26] M. MÉTIVIER, *Semimartingales*, Walter de Gruyter, Berlin, 1982.

[27] M. MINOUX AND J. Y. SERREAULT, *Subgradient optimization and large scale programming: an application to optimum multicommodity network synthesis with security constraints*, Revue RAIRO Recherche Opérationnelle, Dunod, Paris, 15 (1981), pp. 185–203.

[28] E. A. NURMINSKI, *Numerical Methods for Solving Deterministic and Stochastic Minimax Problems*, Naukova Dunka, Kiev, 1979. [In Russian.]

[29] L. ROBBINS AND S. MONRO, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.

[30] R. WETS, *Large scale linear programming techniques in stochastic programming*, in Numerical Techniques for Stochastic Optimization, Y. M. Ermoliev and R. Wets, eds., Springer-Verlag, Berlin, 1988.

[31] A. P. WIERBICKI AND S. KURCYUSZ, *Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space*, SIAM J. Control Optim., 15 (1977), pp. 25–56.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1404–1419, November 1990

007

HAMILTON–JACOBI THEORY FOR OPTIMAL CONTROL PROBLEMS WITH DATA MEASURABLE IN TIME*

R. B. VINTER† AND P. WOLENSKI†

Abstract. Hamilton-Jacobi theory provides necessary and sufficient conditions on minimizing arcs in terms of solutions to the Hamilton-Jacobi equation or inequality. The hypotheses under which such results have previously been obtained typically require the data to be continuous in its time-dependence. The present paper lifts this restriction. The basic hypotheses are Carathéodory-type with measurable time and Lipschitz state dependence, and they incorporate the growth condition of Valadier's existence theory. It is shown that the value function is a solution to the Hamilton-Jacobi equation in an extended sense defined in terms of lower Dini directional derivatives, and that solutions of the related inequality furnish verification functions. Moreover, a characterization of the value function is provided as the pointwise maximum of the family of all verification functions. The methods developed to take account of the measurable time-dependence are based on a "uniform" Lebesgue point theorem for integrably bounded set-valued functions.

Key words. dynamic programming, differential inclusions, nonsmooth analysis

AMS(MOS) subject classifications. 49C05, 49C20

1. Introduction. We consider the following differential inclusion formulation of the optimal control problem (it is labelled (P)):

(P) Minimize $f(x(1))$ over $x(\cdot) \in AC[0,1]$ such that
$\dot{x}(t) \in F(t, x(t))$, a.e. $t \in [0,1]$ and $x(0) = x_0$.

Here $f: \Re^n \to \Re^1$ is a locally Lipschitz continuous function, x_0 is a given point in $\Re^n$, and $F: [0,1] \times \Re^n \rightrightarrows \Re^n$ is a multifunction (or set-valued map). $\dot{x}(t)$ denotes the derivative of $x(\cdot)$.

The value function $V: [0,1] \times \Re^n \to \Re^n$ is defined as follows: for every point $(t, \xi) \in [0,1] \times \Re^n$, $V(t, \xi)$ is the minimal cost of a modified version of problem (P) in which $[t, 1]$ replaces $[0, 1]$ as the underlying time interval and ξ replaces x_0 as the initial condition. Under the hypotheses we shall soon impose, V will be finite everywhere.

Hamilton-Jacobi theory as it bears on problem (P) contains two important elements: verification theorems and properties of the value function. The Hamilton-Jacobi equation (or the related inequality) provides the link between them. On the one hand, a verification theorem (or the Hamilton-Jacobi verification technique) is a sufficient condition for an arc to be minimizing. This is expressed in the terms of "verification functions," which are solutions to the Hamilton-Jacobi equation suitably defined. On the other hand, the value function furnishes a solution to the Hamilton-Jacobi equation in an appropriate sense.

A restrictive feature of early forms of verification theorems (see, e.g., [9]) is the requirement that verification functions be classical C^1 solutions to the Hamilton-Jacobi equation. Yet for many problems we should like to address, no C^1 solutions exist. Recent research has been directed at weakening the requirements on verification functions to the point where, under mild and verifiable hypotheses, conditions are provided which are necessary as well as sufficient for optimality. Nonclassical analysis has had a major role here, since the verification functions we typically encounter are nondifferentiable. A variety of approaches have been followed. For example, methods

* Received by the editors January 25, 1989; accepted for publication (in revised form) January 15, 1990.
† Department of Electrical Engineering, Imperial College, Exhibition Road, London SW7 2BT, United Kingdom.

have been based on the notion of viscosity solutions [15], [8], [17], almost everywhere strict sense solutions [3], [11], Krotov functions [24], Clarke generalized gradients [7], a sequence of solutions to the Hamilton-Jacobi inequality [27], and lower Dini (or contingent) derivatives [10], [4]. See also [20], [21], and [25].

We focus attention on results where the definition of verification functions involves generalized gradients, and that provide necessary and sufficient conditions of optimality. Here it is customary (see, e.g., [7], [10], or [4]) to include among the hypotheses:

(E) F is continuous with respect to the Hausdorff metric.

(We mention, however, that Ishii [14] and Lions and Perthame [16] have established existence and uniqueness of viscosity-type solutions to the Hamilton-Jacobi equation in situations where the data is measurable in time.)

We shall show that if we define a generalized solution to the Hamilton-Jacobi equation to be a function in a certain function space (strictly larger than the space of locally Lipschitz continuous functions) which is a lower Dini solution to the Hamilton-Jacobi equation, then we obtain necessary and sufficient conditions of optimality when hypothesis (E) is relaxed to require merely measurable dependence of F in the time variable. By "lower Dini solution" we mean a function ϕ which satisfies

$$\min_{v\in F(t,\xi)} d^-\phi((t,\xi);(1,v)) \geqq 0 \tag{1.1}$$

on some suitable set. Here we use $d^-\phi(\eta; w)$ to denote the lower Dini derivative of ϕ in the direction w:

$$d^-\phi(\eta; w) := \liminf_{\substack{h\downarrow 0\\ u\to w}} \frac{1}{h}\{\phi(\eta+hu)-\phi(\eta)\}.$$

Our proof techniques accord a prominent role to properties of the set function defined by the reachable set as it evolves in time (cf. [22], [29], [30]). They also make use of a "uniform" extension of previously available results on Lebesgue points of set-valued mappings, originating in the work of Hermes [12].

Specifically, we shall work with the following hypotheses (these will often be referred to as the *basic assumptions* on F):

(H1) For all $(t,\xi)\in[0,1]\times\mathfrak{R}^n$, $F(t,\xi)$ is a nonempty compact set.

(H2) There exists a function $\lambda_1(\cdot)\in L^1[0,1]$ so that for all $t\in[0,1]$, $\xi,\xi'\in\mathfrak{R}^n$, we have

$$\operatorname{dist}_H(F(t,\xi),F(t,\xi')) \leqq \lambda_1(t)|\xi-\xi'|,$$

where dist_H denotes the Hausdorff metric.

(H3) For all $\xi\in\mathfrak{R}^n$, the multifunction $t\rightrightarrows F(t,\xi)$ is measurable, and there exist functions $r(\cdot)$, $\lambda_2(\cdot)\in L^1[0,1]$ so that

$$F(t,\xi)\subseteq(r(t)+\lambda_2(t)|\xi|)B,$$

where B denotes the closed unit ball.

Of particular interest here is the fact that we permit F to be merely measurable as a function of time.

There are two principal reasons for working with these hypotheses. First, it is desirable to have a broad, common framework of hypotheses within which to develop different branches of optimal control, and in particular dynamic programming and the theory of necessary conditions. The point is that, in proving certain results, we need

simultaneously to draw on the theory of dynamic programming and to apply necessary conditions of optimality. Such results will be rather restrictive unless the two bodies of hypotheses under which each is valid overlap to a large extent. (One instance of this is the interpretation of the costate variable in terms of generalized gradients of the value function [26]. Another is proof of verification theorems in the presence of endpoint constraints [7]; here necessary conditions of optimality have a major role.) The theory of first-order necessary conditions has been developed for problems with data measurable in the time variable and Lipschitz continuous in the state variable, and so we would also like to treat such problems in dynamic programming.

Second, problems naturally arise, for example in optimal resource extraction, where the data is discontinuous in the time variable. They typically involve abrupt changes in tariffs or interest rates. The following example is a special case of the hydroelectric power extraction problem in [18], in which the reservoir is assumed to have vertical sides and there is no flow into the reservoir.

Example.

Minimize $-x_1(1)$

subject to

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} \in \left\{ \begin{pmatrix} w(t)x(t)u \\ -u \end{pmatrix} : u \in [0, 1] \right\} \quad \text{a.e. } [0, 1],$$

$$\begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = \begin{pmatrix} 0 \\ \xi \end{pmatrix}.$$

Here $w(\cdot)$, the tariff function, is a given bounded, measurable function. ξ, a given positive number, is the head at the turbine at time $t = 0$.

In this problem $x_2(t)$ is the head at time t, whose rate of decrease is proportional to the flow rate u through the turbine. $x_1(t)$ is the profit which has accrued up to time t. Consequently, the objective is to maximize the profit over the time interval $[0, 1]$. A case of interest is that where the tariff is piecewise constant [18]; the different levels represent charges for consumption of electricity over different periods of the day. In this case, the data is discontinuous in time.

We shall solve this problem, for certain choices of $w(\cdot)$, using a verification function treating data measurable in the time variable.

Results such as those reported in this paper do not supply a methodology for obtaining verification functions. Their main purpose is to establish that if we have obtained a putative minimizer (by finding an extremal, for example), then we can in principle settle the question of whether it is optimal by finding a verification function.

Lower Dini derivatives are by no means new to the optimization and dynamical systems literature. Work on Lyapunov functions by Yorke [31], necessary conditions of optimality due to Ioffe [13], and verification theorems of Frankowska [10] and Berkovitz [4] all make use of them in one way or another, to name but a few instances. However, this paper apparently identifies for the first time the significance of lower Dini derivatives in the treatment of optimal control problems with data measurable in time.

We comment on some implications of our findings, all of which concern extensions of known results to situations where the data is measurable in time. The fact that the value function is a lower Dini solution of the Hamilton-Jacobi inequality is the basis of the proof in [28] that a minimizing arc $x(\cdot)$ for problem (P) and an associated co-extremal $p(\cdot)$ are related according to

$$(h(t), -p(t)) \in \partial V(t, x(t)) \quad \text{a.e. } t \in [0, 1], \tag{1.2}$$

where $h(\cdot) := \max_{v \in F(t,x(t))} p(t) \cdot v$. ("$\partial$" refers to the Clarke subgradient.) Inclusion (1.2) extends the results of [26] to allow measurability in time. Using the exact penalization techniques of [7], we obtain necessary and sufficient conditions of optimality (valid for data measurable in time) for a problem related to (P) in which a right endpoint constraint "$x(1) \in C$" is introduced. A more intricate application of our methods leads to a characterization of *local* minimizers for (P), which involves verification theorems having domain a tube about the arc under consideration (cf. [7]).

We conclude this Introduction with a few definitions and identities. For a time interval $[t_0, t_1] \subset [0, 1]$ and an initial condition $\xi \in \mathfrak{R}^n$, the set $S(t_0, t_1, \xi)$ comprises the solution $x(\cdot) \in AC[t_0, t_1]$ to the differential inclusion

$$\begin{aligned} &\dot{x}(t) \in F(t, x(t)) \quad \text{a.e. } t \in [t_0, t_1], \\ &x(t_0) = \xi. \end{aligned} \tag{1.3}$$

We denote by $S_0(t_0, t_1, \xi)$ the solutions to the associated convexified differential inclusion. That is,

$$S_0(t_0, t_1, \xi) := \{x(\cdot) \in AC[t_0, t_1]: x(t_0) = \xi \text{ and } \dot{x}(t) \in \text{co } F(t, x(t)) \text{ a.e. } t \in [t_0, t_1]\}$$

("co" signifies "convex hull"). The *reachable set* $R(t_0, t_1, \xi)$ (with time interval $[t_0, t_1] \subseteq [0, 1]$ and initial state $\xi \in \mathfrak{R}^n$) is defined by

$$R(t_0, t_1, \xi) := \{x(t_1) \in \mathfrak{R}^n: x(\cdot) \in S(t_0, t_1, \xi)\}.$$

Under hypotheses (H1)-(H3) the reachable set is nonempty (a consequence of Valadier's existence theory [23]) and bounded (see, for example, Lemma 3.1 below). It is evident that

$$V(t, \xi) = \inf \{f(x(1)): x(\cdot) \in S(t, 1, \xi)\}. \tag{1.4}$$

Again when (H1)-(H3) are in force, the Filippov-Wazewski relaxation theorem (see, e.g., [6, p. 117]) is applicable and asserts, in particular, that

$$\text{cl } R(t_0, t_1, \xi) = \{x(t_1): x(\cdot) \in S_0(t_0, t_1, \xi)\}.$$

Since the cost function f is continuous, this identity yields another characterization of the value function

$$V(t, \xi) = \min \{f(x(1)): x(\cdot) \in S_0(t, 1, \xi)\},$$

where the notation "min" indicates that a minimizing arc $x(\cdot)$ exists.

It will be convenient in the proofs below to use only one function $\lambda(\cdot) \in L^1$. Define $\lambda(t) := \max \{\lambda_1(t), \lambda_2(t)\}$. Note that if (H2) and (H3) hold, then they remain satisfied for $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ replaced by $\lambda(\cdot)$.

2. The main results. Our purpose is to demonstrate the existence of verification functions characterizing optimal arcs under the mild hypotheses (H1)-(H3) (the "basic assumptions"). The choice of function space for the class of verification functions under consideration is a delicate matter. It is dictated by the regularity properties of the value function, which is, after all, the natural candidate for a verification function. These properties cannot be described with adequate precision by appealing to the standard spaces, for example, those comprising Lipschitz continuous or continuous functions. It is well known that the value function is locally Lipschitz continuous in the state variable and uniformly so with respect to the time variable. It will be shown below that the value function is also, in some sense, "uniformly" absolutely continuous in the time variable, as the state variable ranges over an arbitrary compact subset. More precisely, we shall see that the value function meets the requirements of the following definition.

DEFINITION 2.1. Suppose $\phi:[0,1]\times\Re^n\to\Re^1$. We say that $\phi(\cdot,\xi)$ is absolutely continuous independent of ξ in a compact set if, for all $\varepsilon>0$ and $K\subseteq\Re^n$ compact, there exists $\delta>0$ so that for any finite collection $[a_1,b_1],\cdots,[a_m,b_m]$ of disjoint subintervals of $[0,1]$ satisfying $\sum_{j=1}^m(b_j-a_j)<\delta$, we have

$$\sum_{j=1}^m \sup_{\xi\in K}|\phi(b_j,\xi)-\phi(a_j,\xi)|<\varepsilon.$$

These conditions satisfied by the value function—local Lipschitz continuity in ξ and absolute continuity in t in some uniform sense—are less restrictive than local Lipschitz continuity jointly in ξ and t (see the remark in § 6). They are built into the following definition of a verification function which also must satisfy the Hamilton-Jacobi inequality (in a lower Dini sense) and the appropriate boundary conditions.

DEFINITION 2.2. Suppose $\phi:[0,1]\times\Re^n\to\Re^1$. Then ϕ is called a verification function (for the problem (P)) if

(a) $\phi(\cdot,\xi)$ is absolutely continuous independent of ξ in a compact set,

(b) $\phi(t,\cdot)$ is locally Lipschitz with Lipschitz constant independent of $t\in[0,1]$,

(c) $\phi(1,\xi)=f(\xi)$ for all ξ, and

(d) there exists $J\subseteq[0,1)$ of measure zero so that for all $t\in[0,1)\backslash J$ and $\xi\in\Re^n$, we have

$$\min_{v\in F(t,\xi)} d^-\phi((t,\xi);(1,v))\geqq 0. \tag{2.1}$$

Our first result asserts that the value function V is indeed a verification function, and that V is identifiable as the maximal element in the set of verification functions. Furthermore, V is a special verification function for which a stronger Hamilton-Jacobi equality replaces the inequality (2.1).

THEOREM 2.3. *Suppose F satisfies the basic assumptions. Then V is a verification function such that for each $(t,\xi)\in[0,1]\times\Re^n$,*

$$V(t,\xi)=\max\{\phi(t,\xi):\ \phi \text{ is a verification function}\}. \tag{2.2}$$

Moreover, the null set J in Definition 2.2(d) can be chosen so that V satisfies the Hamilton-Jacobi equation in the following sense. For $t\in[0,1)\backslash J$ and $\xi\in\Re^n$, we have

$$\min_{v\in\operatorname{co}F(t,\xi)} d^-V((t,\xi):(1,v))=0. \tag{2.3}$$

Gonzales draws attention to the fact that the value function is the maximal element in the class of verification functions (see equation (2.2)) in [11], which treats problems with smooth data. Such characterizations are implicit in earlier literature however (see references in [27].)

And now we give necessary and sufficient conditions of optimality, expressed in terms of verification functions.

THEOREM 2.4 (the Hamilton-Jacobi verification technique). *Suppose F satisfies the basic assumptions. Then $x(\cdot)\in S(0,1,x_0)$ is an optimal solution to (1.1) if and only if there exists a verification function ϕ so that $\phi(0,x_0)=f(x(1))$.*

Suppose it is known that the value function is locally Lipschitz continuous jointly in the time and state variables. This will be the case, for example, when the functions r and λ_2 of hypothesis (H3) are essentially bounded. Then the assertions of Theorems 2.3 and 2.4 remain true when a simpler definition of verification function is adopted: We may replace (a) and (b) on ϕ in Definition 2.2 by "ϕ is locally Lipschitz continuous in (t,ξ) jointly."

If F is continuous in both variables, then the null set J in Theorem 2.3 can be taken to be the empty set. Under the continuity assumptions, Frankowska [10] has recently shown that equation (2.3) holds at all $(t, \xi) \in [0, 1) \times \mathfrak{R}^n$. Her proof makes use of a convergence property of reachable sets; an important ingredient in our analysis is the demonstration that this property is preserved when we relax the hypotheses to permit, among other things, measurable time-dependence (see Lemma 4.3 below). This convergence property (in the case of continuous data) features also in the work of Roxin [22].

We conclude the section with a brief analysis of the hydroelectric power extraction problem example of § 1. Consider the case when the tariff function is monotone decreasing. Here the minimizing trajectory $(\bar{x}_1, \bar{x}_2)$ is given by

$$\bar{x}_1(t) = \int_0^t w(s)[\xi - s]^+ \, ds,$$
$$\bar{x}_2(t) = [\xi - t]^+,$$

in which $a^+ := \max\{a, 0\}$. The interpretation of this trajectory is that maximum profits are obtained (for a monotone decreasing tariff function) by operating the turbine at maximum flow until the reservoir is emptied.

Optimality of this trajectory can be confirmed by application of Theorem 2.4 in which we adopt as verification function the locally Lipschitz continuous function

$$W(t, x) = -x_1 - \int_t^1 w(s)[x_2 + t - s]^+ \, ds.$$

This function is a lower Dini solution of the Hamilton-Jacobi equation at all points in $S \times \mathfrak{R}^2$, where S is the subset of $[0, 1)$ on which w is continuous.

3. Properties of reachable sets. In this section, we show that (a) and (b) of Definition 2.2 hold for $\phi = V$. The proofs of these facts rely on standard properties of reachable sets and the observation (1.4). Throughout this section, it is assumed that F satisfies the basic assumptions.

LEMMA 3.1. *Let $K \subseteq \mathfrak{R}^n$ be compact. Then there exist $\gamma(\cdot) \in L^1[0, 1]$ and a constant $k_1 > 0$ (both depending on K) so that for all $\xi \in K$ and $0 \leqq t_0 < t_1 \leqq 1$, we have that $\eta \in R(t_0, t_1, \xi)$ implies*

(i) $|\eta - \xi| \leqq \int_{t_0}^{t_1} \gamma(s) \, ds$, *and*

(ii) $|\eta| \leqq k_1$.

Proof. (i) Let $\xi \in K$ and $0 \leqq t_0 < t_1 \leqq 1$. Then for each $x(\cdot) \in S(t_0, t_1, \xi)$ and $t \in [t_0, t_1]$, we have

$$\begin{aligned} |x(t) - \xi| &\leqq \int_{t_0}^t |\dot{x}(s)| \, ds \\ &\leqq \int_{t_0}^t (r(s) + \lambda(s)|x(s)|) \, ds \quad \text{(by (H3))} \\ &\leqq \int_{t_0}^t (r(s) + \lambda(s)|\xi|) \, ds + \int_{t_0}^t \lambda(s)|x(s) - \xi| \, ds. \end{aligned}$$

An application of Gronwall's inequality (see, e.g., [2, p. 119]) gives

$$\begin{aligned} |x(t_1) - \xi| &\leqq \int_{t_0}^{t_1} (r(s) + \lambda(s)|\xi|) \, ds + \int_{t_0}^{t_1} \lambda(s) \\ &\quad \times \left(\int_{t_0}^s r(s') + \lambda(s')|\xi| \, ds' \right) \exp\left(\int_s^{t_1} \lambda(s') \, ds' \right) ds \\ &\leqq [1 + \|\lambda\|_1 \exp \|\lambda\|_1] \int_{t_0}^{t_1} (r(s) + \lambda(s)|K|) \, ds, \end{aligned}$$

where $|K| := \max\{|\eta|: \eta \in K\}$. Therefore (i) holds for

$$\gamma(s) := (1 + \|\lambda\|_1 \exp \|\lambda\|_1)(r(s) + \lambda(s)|K|).$$

(ii) Let $k_1 = \|\gamma\|_1 + |K|$. Then (ii) follows immediately from (i). □

LEMMA 3.2. *For fixed* $0 \leqq t_0 < t_1 \leqq 1$, *the multifunction* $\xi \rightrightarrows \text{cl } R(t_0, t_1, \xi)$ *is Lipschitz (with respect to the Hausdorff metric) of order* $\exp(\int_{t_0}^{t_1} \lambda(s)\, ds)$.

Proof (see Aubin and Cellina [2, pp. 120–123]). Although the theorem in this reference does not explicitly allow for measurable dependence of $F(\cdot, \xi)$, the proof can easily be adapted to incorporate this more general feature. □

LEMMA 3.3. *Let* $\varepsilon > 0$ *and* $K \subseteq \mathfrak{R}^n$ *be compact. Then there exists* $\delta = \delta(\varepsilon, K) > 0$ *so that for any finite collection* $[a_1, b_1], \cdots, [a_m, b_m]$ *of disjoint subintervals of* $[0, 1]$ *satisfying* $\sum_{j=1}^{m} (b_j - a_j) < \delta$, *we have*

$$\sum_{j=1}^{m} \sup_{\xi \in K} \text{dist}_H\,(\text{cl } R(b_j, 1, \xi), \text{cl } R(a_j, 1, \xi)) < \varepsilon. \tag{3.1}$$

Proof. Let k_1 be as in Lemma 3.1, and set $k_2 = \exp\{\int_0^1 \lambda(t)\, dt\}$. Define $\alpha: [0, 1] \to \mathfrak{R}^1$ by $\alpha(t) = k_2 \int_0^t (r(s) + \lambda(s)k_1)\, ds$. Because $\alpha(\cdot) \in AC[0, 1]$, to prove the lemma it suffices to show that for all $0 \leqq t_0 < t_1 \leqq 1$ and $\xi \in K$, we have $\text{dist}_H\,(\text{cl } R(t_0, 1, \xi), \text{cl } R(t_1, 1, \xi)) \leqq \alpha(t_1) - \alpha(t_0)$.

Fix $0 \leqq t_0 < t_1 \leqq 1$ and $\xi \in K$, and suppose $x(\cdot) \in S_0(t_0, 1, \xi)$. Obviously, $x(\cdot)$ restricted to $[t_1, 1]$ is an element of $S_0(t_1, 1, x(t_1))$. Hence by Lemma 3.2, there exists $y(\cdot) \in S_0(t_1, 1, \xi)$ so that

$$\begin{aligned}
|x(1) - y(1)| &\leqq k_2 |x(t_1) - \xi| \\
&\leqq k_2 \int_{t_0}^{t_1} |\dot{x}(s)|\, ds \\
&\leqq k_2 \int_{t_0}^{t_1} (r(s) + \lambda(s)k_1)\, ds \quad \text{(by (H3) and Lemma 3.1(ii))} \\
&= \alpha(t_1) - \alpha(t_0).
\end{aligned}$$

Thus we have shown that $\text{cl } R(t_0, 1, \xi) \subseteq \text{cl } R(t_1, 1, \xi) + (\alpha(t_1) - \alpha(t_0))B$.

Now suppose $y(\cdot) \in S_0(t_1, 1, \xi)$. Pick any $x_1(\cdot) \in S_0(t_0, t_1, \xi)$. By Lemma 3.2, there exists $x_2(\cdot) \in S_0(t_1, 1, x_1(t_1))$ so that

$$\begin{aligned}
|x_2(1) - y(1)| &\leqq k_2 |x_2(t_1) - y(t_1)| \\
&= k_2 |x_1(t_1) - \xi| \\
&\leqq \alpha(t_1) - \alpha(t_0) \quad \text{(as above)}.
\end{aligned}$$

Since $x_2(1) \in \text{cl } R(t_0, 1, \xi)$, we now have shown $\text{cl } R(t_1, 1, \xi) \subseteq \text{cl } R(t_0, 1, \xi) + (\alpha(t_1) - \alpha(t_0))B$. Combining this with the above, we conclude that $\text{dist}_H\,(\text{cl } R(t_0, 1, \xi), \text{cl } R(t_1, 1, \xi)) \leqq \alpha(t_1) - \alpha(t_0)$. □

We are now ready to show that properties (a) and (b) of Definition 2.2 hold for $\phi = V$.

Proof of (a). Let $\varepsilon > 0$ and $K \subseteq \mathfrak{R}^n$ be compact. Let k_1 be the constant of Lemma 3.1 and l the Lipschitz constant of f on $k_1 B$. Fix $\xi \in K$ and $0 \leqq t_0 < t_1 \leqq 1$. Observe from (1.4) that

$$|V(t_0, \xi) - V(t_1, \xi)| \leqq l\, \text{dist}_H\,(\text{cl } R(t_0, 1, \xi), \text{cl } R(t_1, 1, \xi)) \tag{3.2}$$

holds. Now choose δ as in Lemma 3.3 for ε replaced by ε/l in (3.1). Then for any finite collection $[a_1, b_1], \cdots, [a_m, b_m]$ of disjoint subintervals of $[0, 1]$ satisfying $\sum_{j=1}^{m} (b_j - a_j) < \delta$, we have

$$\sum_{j=1}^{m} \sup_{\xi \in K} |V(b_j, 1, \xi) - V(a_j, 1, \xi)|$$

$$\leqq l \sum_{j=1}^{m} \sup_{\xi \in K} \operatorname{dist}_H (\operatorname{cl} R(b_j, 1, \xi), \operatorname{cl} R(a_j, 1, \xi)) \quad \text{(by (3.2))}$$

$$< \varepsilon \quad \text{(by (3.1))}.$$

Hence $V(\cdot, \xi)$ is absolutely continuous independent of ξ in a compact set. □

Proof of (b). Let $K \subseteq \mathfrak{R}^n$ be compact, and let l be as in the proof of (a). The analogue of (3.2) with varying state variables is: for $\xi, \xi' \in K$ and $t \in [0, 1]$, we have

$$|V(t, \xi) - V(t, \xi')| \leqq l \operatorname{dist}_H (\operatorname{cl} R(t, 1, \xi), \operatorname{cl} R(t, 1, \xi')).$$

From this it follows via Lemma 3.2 that for all $\xi, \xi' \in K$ and $t \in [0, 1]$, we have

$$|V(t, \xi) - V(t, \xi')| \leqq l \exp\left(\int_0^1 \lambda(s)\, ds\right) |\xi - \xi'|.$$

This proves (b). □

4. Lebesgue points of integrably bounded multifunctions. Before proceeding, it is convenient to review some notions of set convergence. For $\xi \in \mathfrak{R}^n$ and $K \subseteq \mathfrak{R}^n$, define the distance from ξ to K by $\operatorname{dist}(\xi, K) = \inf\{|\xi - \eta| : \eta \in K\}$. If $\{K(t)\}_{0 < t \leqq t_0}$ is a collection of subsets of $\mathfrak{R}^n$ parameterized by $t \in (0, t_0]$, then the (Kuratowski) lim inf and lim sup of $K(\cdot)$ as $t \downarrow 0$ are defined by

$$\liminf_{t \downarrow 0} K(t) = \{\eta : \limsup_{t \downarrow 0} \operatorname{dist}(\eta, K(t)) = 0\},$$

$$\limsup_{t \downarrow 0} K(t) = \{\eta : \liminf_{t \downarrow 0} \operatorname{dist}(\eta, K(t)) = 0\}.$$

Of course we always have $\liminf_{t \downarrow 0} K(t) \subseteq \limsup_{t \downarrow 0} K(t)$. If, in fact, we have equality, then we say that the limit exists and write $\lim_{t \downarrow 0} K(t)$ for the common value. Note that both $\liminf_{t \downarrow 0} K(t)$ and $\limsup_{t \downarrow 0} K(t)$ are always closed sets. Also note if each $K(t)$ is convex, then $\liminf_{t \downarrow 0} K(t)$ is also convex.

If $K : [0, 1] \rightrightarrows \mathfrak{R}^n$ is any closed-valued integrably bounded multifunction, the integral of K over $[t_0, t_1] \subseteq [0, 1]$ is the subset of $\mathfrak{R}^n$ given by

$$\int_{t_0}^{t_1} K(s)\, ds := \left\{ \int_{t_0}^{t_1} g(s)\, ds : g(\cdot) \text{ is measurable on } [t_0, t_1] \text{ and } g(s) \in K(s) \text{ a.e. } s \in [t_0, t_1] \right\}.$$

By Auman's theorem (see, e.g., [6, p. 112]), $\int_{t_0}^{t_1} K(s)\, ds$ is always closed and convex.

Recall that the set of right Lebesgue points $L(g)$ of a function $g \in L^1[0, 1]$ is

$$L(g) := \left\{ t : \lim_{h \downarrow 0} \frac{1}{h} \int_t^{t+h} g(s)\, ds \text{ exists, is finite, and equals } g(t) \right\}.$$

It is well known that the set $L(g)$ has full measure in $[0, 1]$. Such results have been extended by Hermes [12] (see also [1], [5]) to situations where a multifunction replaces

the function g. We now prove a further extension, which asserts (under appropriate hypotheses) that the set of "Lebesgue points" of a multifunction, depending on a parameter, has full measure in some *uniform* sense. The hypotheses in question are those of § 1, and the parameter is the state vector x.

PROPOSITION 4.1. *Suppose* $F:[0,1]\times\Re^n \rightrightarrows \Re^n$ *satisfies* (H1)-(H3). *Then there exists* $J\subseteq[0,1)$ *of measure zero so that for all* $t\in[0,1)\backslash J$ *and* $\xi\in\Re^n$, *we have* $\lim_{h\downarrow 0}(1/h)\int_0^h F(t+s,\xi)\,ds$ *exists and equals* co $F(t,\xi)$.

Proof. Our starting point is the fact that for fixed $\xi\in\Re^n$, a null set $J(\xi)\subseteq[0,1)$ exists so that co $F(t,\xi)\subseteq \liminf_{t\downarrow 0}(1/h)\int_0^h F(t+s,\xi)\,ds$ for all $t\in[0,1)\backslash J(\xi)$.

This property, along with other information, is contained in Theorem 4.3 of [1]; it is a consequence of adopting a Castaing representation for the measurable multifunction $F(\cdot)$ (see, e.g., [19]). Fix $\xi\in\Re^n$. Let $\{g_j(t)\}_{j=1}^\infty$ be a Castaing representation. This means $g_j(\cdot)$ is measurable and $F(t,\xi)=\mathrm{cl}\,\{g_j(t)\}_{j=1}^\infty$ for all $t\in[0,1]$. By (H3), $g_j(\cdot)\in L^1[0,1]$ for each j, thus $J(\xi):=(\cup_{j=1}^\infty [0,1)\backslash L(g_j))$ has measure zero. If $t\in[0,1)\backslash J(\xi)$, then for each j we have

$$g_j(t)=\lim_{h\downarrow 0}\frac{1}{h}\int_0^h g_j(t+s)\,ds\in\lim_{h\downarrow 0}\inf\frac{1}{h}\int_0^h F(t+s,\xi)\,ds. \tag{4.1}$$

Since the right side of (4.1) is a closed convex set, by taking the closed convex hull over j on the left side of (4.1), we conclude that $t\in[0,1)\backslash J(\xi)$ implies

$$\text{co } F(t,\xi)\subseteq\lim_{h\downarrow 0}\inf\frac{1}{h}\int_0^h F(t+s,\xi)\,ds. \tag{4.2}$$

Now let $\{\xi_i\}_{i=1}^\infty$ be a countable dense subset of $\Re^n$, and let $J(\xi_i)$ be chosen as above so that (4.2) holds for $\xi=\xi_i$. Define $J_0=([0,1)\backslash L(\lambda))\cup(\cup_{i=1}^\infty J(\xi_i))$, where $\lambda(\cdot)\in L^1[0,1]$ is given by (H2). Then J_0 has measure zero. Fix $t\in[0,1)\backslash J_0$ and $\xi\in\Re^n$. Let $\{h_k\}_{k=1}^\infty$ be an arbitrary sequence with $h_k\downarrow 0$, and let $\varepsilon>0$ and $v\in F(t,\xi)$ also be arbitrary. There exists i_0 so that $|\xi-\xi_{i_0}|<\varepsilon$. By (H2) there exists $v_0\in F(t,\xi_{i_0})$ so that $|v-v_0|<\lambda(t)\varepsilon$. Since $t\in[0,1)\backslash J(\xi_{i_0})$, by (4.2) there exists $\{u_k(\cdot)\}_{k=1}^\infty$ so that

$$\begin{aligned} &u_k(s)\in F(t+s,\xi_{i_0}) \quad \text{a.e. } s\in[0,h_k], \quad \text{and}\\ &\frac{1}{h_k}\int_0^{h_k}u_k(s)\,ds\to v_0 \quad \text{as } k\to\infty. \end{aligned} \tag{4.3}$$

Again by (H2), the set $P_k(s):=F(t+s,\xi)\cap(u_k(s)+\lambda(t+s)|\xi-\xi_{i_0}|B)$ is nonempty almost everywhere on $[0,h_k]$, and by Theorem 1M of [19] the multifunction $P_k(\cdot)$ is measurable on $[0,h_k]$. Let $v_k(\cdot)$ be a measurable selection of $P_k(\cdot)$ (see [19, Cor. 1C]). Then

$$\begin{aligned} \left|\frac{1}{h_k}\int_0^{h_k}v_k(s)\,ds-v\right| &\leqq \frac{1}{h_k}\int_0^{h_k}|v_k(s)-u_k(s)|\,ds+\left|\frac{1}{h_k}\int_0^{h_k}u_k(s)\,ds-v_0\right|+|v_0-v|\\ &\leqq \frac{\varepsilon}{h_k}\int_0^{h_k}\lambda(t+s)\,ds+\left|\frac{1}{h_k}\int_0^{h_k}u_k(s)\,ds-v_0\right|+\lambda(t)\varepsilon. \end{aligned} \tag{4.4}$$

Now let $k\to\infty$ and apply (4.3); the middle term on the right side in (4.4) approaches zero. Since $t\in L(\lambda)$, and ε is arbitrary, we conclude from (4.4) that $(1/h_k)\int_0^{h_k}v_k(s)\,ds\to v$. Because

$$\frac{1}{h_k}\int_0^{h_k}v_k(s)\,ds\in\frac{1}{h_k}\int_0^{h_k}F(t+s,\xi)\,ds$$

for all k, and $\{h_k\}$ and $v \in F(t, \xi)$ are arbitrary, we arrive at $\operatorname{co} F(t, \xi) \subseteq \liminf_{h \downarrow 0} (1/h) \int_0^h F(t+s, \xi)\, ds$.

To finish the proof, we show there exists a null set J_1 so that $t \in [0, 1) \backslash J_1$ and $\xi \in \mathfrak{R}^n$ imply $\limsup_{h \downarrow 0} (1/h) \int_0^h F(t+s, \xi)\, ds \subseteq \operatorname{co} F(t, \xi)$. First, fix sequences $\{\xi_i\}_{i=1}^{\infty}$ and $\{p_j\}_{j=1}^{\infty}$ that are both dense in $\mathfrak{R}^n$. For $(t, \xi, p) \in [0, 1] \times \mathfrak{R}^n \times \mathfrak{R}^n$, define the Hamiltonian by $H(t, \xi, p) := \sup_{v \in F(t, \xi)} \langle v, p \rangle$. Note that by (H3), for each ξ and p, the function $t \to H(t, \xi, p)$ is an element of $L^1[0, 1]$. Define $J_1 := [0, 1) \backslash \{L(\lambda) \cap (\cap_{i,j} L(H(\cdot, \xi_i, p_j)))\}$, where, as earlier, $L(g)$ denotes the right Lebesgue points of $g \in L^1[0, 1]$. Obviously J_1 has measure zero. We remark that $v \in \operatorname{co} F(t, \xi)$ if and only if $\langle v, p_j \rangle \leqq H(t, \xi, p_j)$ for all $j = 1, 2, \cdots$; this is a result from elementary convexity theory. Also note that a consequence of assumption (H2) is $\xi \to H(t, \xi, p)$ is continuous for fixed t and p.

Now fix $t \in [0, 1) \backslash J_1$ and $\xi \in \mathfrak{R}^n$. Let $v \in \limsup_{t \downarrow 0} (1/h) \int_0^h F(t+s, \xi)\, ds$. By definition, there exists $h_k \downarrow 0$ and $g_k(\cdot) \in L^1[0, h_k]$ with $g_k(s) \in F(t+s, \xi)$ almost everywhere $s \in [0, h_k]$ so that

$$\frac{1}{h_k} \int_0^{h_k} g_k(s)\, ds \to v \quad \text{as } k \to \infty.$$

For each i and j, we have

$$\begin{aligned} \langle v, p_j \rangle &= \lim_{k \to \infty} \frac{1}{h_k} \int_0^{h_k} \langle g_k(s), p_j \rangle\, ds \\ &\leqq \limsup_{k \to \infty} \frac{1}{h_k} \int_0^{h_k} \{H(t+s, \xi_i, p_j) + \lambda(t+s) |\xi - \xi_i|\}\, ds \quad \text{(by (H2))} \\ &= H(t, \xi_i, p_j) + \lambda(t) |\xi - \xi_i|. \end{aligned} \tag{4.5}$$

The last equality holds because t is a right Lebesgue point of $H(\cdot, \xi_i, p_j)$ and $\lambda(\cdot)$. By considering points in $\{\xi_i\}$ converging to ξ, we deduce from (4.5) that $\langle v, p_j \rangle \leqq H(t, \xi, p_j)$ for all j. Hence the desired result $v \in \operatorname{co} F(t, \xi)$ follows.

We have shown that the conclusion of the proposition holds for $J := J_0 \cup J_1$. □

The next proposition is a widely quoted result by Filippov. It will play a major role in the proof of Lemma 4.3.

PROPOSITION 4.2 (Filippov). *Suppose* $F: [0, 1] \times \mathfrak{R}^n \rightrightarrows \mathfrak{R}^n$ *satisfies* (H1)-(H3), *and let* $\xi \in \mathfrak{R}^n$. *Then there exists a constant* k_3 *so that for all* $0 \leqq t_0 < t_1 \leqq 1$ *and* $y(\cdot) \in AC[t_0, t_1]$, *there is a function* $x(\cdot) \in S(t_0, t_1, y(t_0))$ *satisfying*

$$|x(t_1) - y(t_1)| \leqq k_3 \rho(y),$$

where $\rho(y) := \int_{t_0}^{t_1} \operatorname{dist}(\dot{y}(s), F(s, y(s)))\, ds$.

Proof. See, for example, Theorem 3.1.6 of [6].

LEMMA 4.3. *Suppose* $F: [0, 1] \times \mathfrak{R}^n \rightrightarrows \mathfrak{R}^n$ *satisfies* (H1)-(H3). *Then there exists a set* $J \subseteq [0, 1)$ *of measure zero so that for all* $t \notin J$ *and* $\xi \in \mathfrak{R}^n$, *we have that* $\lim_{h \downarrow 0} (1/h) \times (R(t+h, t, \xi) - \xi)$ *exists and equals* $\operatorname{co} F(t, \xi)$.

Proof. Let J be the null set of Proposition 4.1, and fix $t \in [0, 1) \backslash J$ and $\xi \in \mathfrak{R}^n$. In view of Proposition 4.1, it suffices to show the following.

Let $\{h_k\}$ be an arbitrary sequence of positive numbers with $h_k \downarrow 0$, and let $v \in \mathfrak{R}^n$. Then there exists a sequence of functions $g_k(\cdot) \in L^1[0, h_k]$, $k = 1, 2, \cdots$, with $g_k(s) \in F(t+s, \xi)$ almost everywhere $s \in [0, h_k]$ and such that

$$\frac{1}{h_k} \int_0^{h_k} g_k(s)\, ds \to v \quad \text{as } k \to \infty \tag{4.6}$$

if and only if there exists a sequence $x_k(\cdot) \in S(t, t+h_k, \xi)$, $k = 1, 2, \cdots$, such that

$$\frac{1}{h_k}(x_k(t+h_k) - \xi) \to v \quad \text{as } k \to \infty. \tag{4.7}$$

So now suppose $g_k(\cdot)$ are chosen so that (4.6) holds. We next show $x_k(\cdot)$ exist so that (4.7) holds. Let $y_k(s) = \xi + \int_0^s g_k(s')\, ds'$. Then

$$\begin{aligned}
\rho(y_k) &= \int_0^{h_k} \operatorname{dist}(g_k(s), F(t+s, y_k(s)))\, ds \\
&\leqq \int_0^{h_k} \operatorname{dist}_H(F(t+s, \xi), F(t+s, y_k(s)))\, ds \\
&\leqq \int_0^{h_k} \lambda(t+s)|y_k(s) - \xi|\, ds \quad \text{(by (H2))} \\
&\leqq \delta_k \int_0^{h_k} \lambda(t+s)\, ds,
\end{aligned} \tag{4.8}$$

where $\delta_k := \sup_{0 \leqq s \leqq h_k} |y_k(s) - \xi|$. From Lemma 3.1(i), it follows that $\delta_k \to 0$ as $k \to \infty$. From Proposition 4.2, for each k there exists $x_k(\cdot) \in S(t, t+h_k, \xi)$ so that

$$\begin{aligned}
|x_k(t+h_k) - y_k(h_k)| &\leqq k_3 \rho(y_k) \\
&\leqq k_3 \delta_k \int_0^{h_k} \lambda(t+s)\, ds \quad \text{(by (4.8))}.
\end{aligned} \tag{4.9}$$

Therefore, since $y_k(h_k) - \xi = \int_0^{h_k} g_k(s)\, ds$, we have

$$\begin{aligned}
\left| \frac{x_k(t+h_k) - \xi}{h_k} - v \right| &\leqq \frac{1}{h_k} |x_k(t+h_k) - y_k(h_k)| + \left| \frac{1}{h_k} \int_0^{h_k} g_k(s)\, ds - v \right| \\
&\to 0 \quad \text{as } k \to \infty \quad \text{(by (4.9) and } t \in L(\lambda), \text{ and by (4.6))}.
\end{aligned}$$

Hence (4.7) holds.

Conversely, suppose we are given $x_k(\cdot)$ so that (4.7) holds. For $s \in [0, h_k]$, let $P_k(s) := \{u \in F(t+s, \xi): |u - \dot{x}_k(t+s)| = \operatorname{dist}(\dot{x}_k(t+s), F(t+s, \xi))\}$. Then $P_k(\cdot)$ is a measurable multifunction [see 19, Cor. 1Q] which is nonempty almost everywhere on $[0, h_k]$ by (H2). Hence there exist measurable selections $g_k(\cdot)$ of $P_k(\cdot)$ for each $k = 1, 2, \cdots$. Observe that (H2) implies $|g_k(s) - \dot{x}_k(t+s)| \leqq \lambda(t+s)|x_k(t+s) - \xi|$. Therefore

$$\begin{aligned}
\left| \frac{1}{h_k} \int_0^{h_k} g_k(s)\, ds - v \right| &\leqq \frac{1}{h_k} \int_0^{h_k} |g_k(s) - \dot{x}_k(t+s)|\, ds + \left| \frac{1}{h_k} \int_0^{h_k} \dot{x}_k(t+s)\, ds - v \right| \\
&\leqq \sup_{0 \leqq s \leqq h_k} |x_k(t+s) - \xi| \frac{1}{h_k} \int_0^{h_k} \lambda(t+s)\, ds \\
&\quad + \left| \frac{1}{h_k}(x_k(t+h_k) - x_k(t)) - v \right| \\
&\to 0 \quad \text{as } k \to \infty \quad \text{(by Lemma 3.1(i) and } t \in L(\lambda), \text{ and by (4.7))}.
\end{aligned}$$

Therefore (4.6) holds, and the proof of the lemma is complete. □

5. The Hamilton–Jacobi equation with lower Dini derivatives. We are now ready to show that V is a verification function and satisfies equation (2.3). It was established

in § 3 that the value function satisfies (a) and (b) in Definition 2.2; part (c) is trivial; part (d) is an immediate consequence of Proposition 5.3 below, which in effect says that V is a verification function for the problem (P) with F replaced by co F. We first supply two simple lemmas.

LEMMA 5.1. *Suppose* $\phi:[0,1]\times\mathfrak{R}^n\to\mathfrak{R}^1$ *satisfies* (b) *in Definition* 2.2. *Then for* $(t,\xi)\in[0,1)\times\mathfrak{R}^n$ *and* $v\in\mathfrak{R}^n$, *the Dini derivative reduces to*

$$d^-\phi((t,\xi);(1,v))=\liminf_{h\downarrow 0}\frac{1}{h}\{\phi(t+h,\xi+hv)-\phi(t,\xi)\}. \tag{5.1}$$

Proof. By definition,

$$d^-\phi((t,\xi);(1,v))=\liminf_{\substack{h\downarrow 0\\(s,u)\to(1,v)}}\frac{1}{h}\{\phi(t+hs,\xi+hu)-\phi(t,\xi)\}. \tag{5.2}$$

Property (b) says that there exists a Lipschitz constant l so that for all (s,u) near $(1,v)$, we have

$$\left|\phi\left(t+h,\xi+h\frac{u}{s}\right)-\phi(t+h,\xi+hv)\right|\leqq lh\left|\frac{u}{s}-v\right|.$$

Consequently the lim inf's in (5.1) and (5.2) coincide. □

LEMMA 5.2. *Suppose* $\phi:[0,1]\times\mathfrak{R}^n\to\mathfrak{R}^1$ *satisfies* (b) *in Definition* 2.2. *Let* $(t,\xi)\in[0,1)\times\mathfrak{R}^n$. *Suppose further that we are given numbers* $h_i\downarrow 0$ *as* $i\to\infty$ *and functions* $x_i(\cdot)\in AC[t,t+h_i]$ *with* $x_i(t)=\xi$ *that satisfy*

$$\frac{x_i(t+h_i)-\xi}{h_i}\to v\in\mathfrak{R}^n \quad \text{as } i\to\infty. \tag{5.3}$$

Then

$$\liminf_{i\to\infty}\frac{1}{h_i}\{\phi(t+h_i,x_i(t+h_i))-\phi(t,\xi)\}=\liminf_{i\to\infty}\frac{1}{h_i}\{\phi(t+h_i,\xi+h_iv)-\phi(t,\xi)\}. \tag{5.4}$$

Proof. Property (b) implies that there exists l so that for all large i, we have

$$|\phi(t+h_i,x_i(t+h_i))-\phi(t+h_i,\xi+h_iv)|\leqq lh_i\left|\frac{x_i(t+h_i)-\xi}{h_i}-v\right|. \tag{5.5}$$

It now follows immediately from (5.3) and (5.5) that (5.4) holds. □

PROPOSITION 5.3. *Suppose* F *satisfies the basic assumptions and* f *is locally Lipschitz continuous. Let* J *be as in Lemma* 4.3. *Then for all* $t\in[0,1)\backslash J$ *and* $\xi\in\mathfrak{R}^n$, *we have*

$$\min_{v\in\operatorname{co}F(t,\xi)}d^-V((t,\xi);(1,v))\geqq 0.$$

Proof. Fix $t\in[0,1)\backslash J$ and $\xi\in\mathfrak{R}^n$. Let $h_i\downarrow 0$ as $i\to\infty$ and let $v\in\operatorname{co}F(t,\xi)$ be arbitrary. Since Lemma 4.3 in particular says that $v\in\liminf_{h\downarrow 0}(1/h)\{R(t,t+h,\xi)-\xi\}$, there exists $x_i(\cdot)\in S(t,t+h_i,\xi)$ so that

$$\frac{x_i(t+h_i)-\xi}{h_i}\to v \quad \text{as } i\to\infty.$$

By the principle of optimality, for each i we have

$$V(t+h_i,x_i(t+h_i))-V(t,\xi)\geqq 0.$$

Therefore

$$
\begin{aligned}
0 &\leqq \liminf_{i\to\infty} \frac{1}{h_i}\{V(t+h_i, x_i(t+h_i)) - V(t,\xi)\} \\
&= \liminf_{i\to\infty} \frac{1}{h_i}\{V(t+h_i, \xi+h_i v) - V(t,\xi)\}.
\end{aligned} \tag{5.6}
$$

The last equality is a consequence of Lemma 5.2. Since $\{h_i\}$ arbitrarily approaches zero, we conclude from (5.6) and Lemma 5.1 that

$$d^- V((t,\xi);(1,v)) \geqq 0.$$

Since $v \in \text{co } F(t,\xi)$ was arbitrary, the proof of Proposition 5.3 is complete. □

It has been shown that V is a verification function. We conclude this section with the proof of equation (2.3), which says that in fact equality holds in the conclusion of Proposition 5.3 (for a possibly different null set J).

Proof of equation (2.3). Append to J of Lemma 4.3 the null set $[0,1)\backslash L(r)$, where $r(\cdot) \in L^1[0,1]$ is that given in (H3) (we still call this null set J). We show that (2.3) holds for this new J. Again fix $t \in [0,1)\backslash J$ and $\xi \in \mathfrak{R}^n$. As mentioned in § 1, there exists $x(\cdot) \in S_0(t,1,\xi)$ so that $f(x(1)) = V(t,\xi)$. Let k_1 be as in Lemma 3.1 (with $K = \{\xi\}$). Then

$$
\begin{aligned}
\limsup_{h\downarrow 0} \frac{1}{h}|x(t+h)-\xi| &\leqq \limsup_{h\downarrow 0} \frac{1}{h}\int_0^h |\dot{x}(t+s)|\,ds \\
&\leqq \limsup_{h\downarrow 0} \frac{1}{h}\int_0^h (r(t+s)+\lambda(t+s)k_1)\,ds \quad \text{(by (H3))} \\
&= r(t)+\lambda(t)k_1 < +\infty.
\end{aligned}
$$

Hence there exist a sequence of numbers $h_i \downarrow 0$ as $i \to \infty$ and a vector $v_0 \in \mathfrak{R}^n$ so that

$$\frac{1}{h_i}(x(t+h_i)-\xi) \to v_0 \quad \text{as } i \to \infty.$$

Since $x(t+h_i) \in \text{cl } R(t, t+h_i, \xi)$ for all i, we have that

$$
\begin{aligned}
v_0 &\in \limsup_{h\downarrow 0} \frac{1}{h}(\text{cl } R(t, t+h, \xi) - \xi) \\
&= \text{co } F(t,\xi) \quad \text{(since } t \notin J \text{ and by Lemma 4.3).}
\end{aligned} \tag{5.7}
$$

Hence

$$
\begin{aligned}
\min_{v\in \text{co } F(t,\xi)} d^- V((t,\xi);(1,v)) &\leqq d^- V((t,\xi);(1,v_0)) \quad \text{(by (5.7))} \\
&= \liminf_{h\downarrow 0} \frac{1}{h}\{V(t+h, \xi+hv_0) - V(t,\xi)\} \quad \text{(by Lemma 5.1)} \\
&\leqq \liminf_{i\to\infty} \frac{1}{h_i}\{V(t+h_i, \xi+h_i v_0) - V(t,\xi)\} \\
&= \liminf_{i\to\infty} \frac{1}{h_i}\{V(t+h_i, x(t+h_i)) - V(t,\xi)\} \quad \text{(by Lemma 5.2)} \\
&= 0 \quad \text{(by the principle of optimality).}
\end{aligned}
$$

In combination with Proposition 5.3, this concludes the proof of (2.3). □

6. Properties of verification functions. In this final section, we complete the proof of Theorem 2.3 by showing that (2.2) holds, and then prove Theorem 2.4. The next lemma contains the important fact about verification functions that has motivated Definition 2.1.

LEMMA 6.1. *Suppose* $\phi:[0,1]\times\Re^n\to\Re^1$ *satisfies* (a) *and* (b) *in Definition* 2.2. *Then for* $0\leqq t_0<t_1\leqq 1$ *and all* $y(\cdot)\in AC[t_0,t_1]$, *the function* $t\to\phi(t,y(t))$ *is in* $AC[t_0,t_1]$.

Proof. Without loss of generality, we can take $t_0=0$ and $t_1=1$. Suppose $y(\cdot)\in AC[0,1]$. Let $\varepsilon>0$, set $K:=$ range of $y(\cdot)$. By (b), there exists $l>0$ so that $\xi\to\phi(t,\xi)$ is Lipschitz of order l on K for each $t\in[0,1]$. By (a), there exists $\delta>0$ so that for any finite collection $[a_1,b_1],\cdots,[a_m,b_m]$ of disjoint subintervals of $[0,1]$ satisfying

$$\sum_{j=1}^{m}(b_j-a_j)<\delta, \tag{6.1}$$

we have

$$\sum_{j=1}^{m}\sup_{\xi\in K}|\phi(b_j,\xi)-\phi(a_j,\xi)|<\frac{\varepsilon}{2}. \tag{6.2}$$

Since $y(\cdot)$ is absolutely continuous, we can shrink δ if necessary so that, if (6.1) holds, then so does

$$\sum_{j=1}^{m}|y(b_j)-y(a_j)|<\frac{\varepsilon}{2l}. \tag{6.3}$$

Take a finite collection $[a_1,b_1],\cdots,[a_m,b_m]$ of disjoint subintervals of $[0,1]$ satisfying (6.1). Then

$$\begin{aligned}
&\sum_{j=1}^{m}|\phi(b_j,y(b_j))-\phi(a_j,y(a_j))|\\
&\quad\leqq\sum_{j=1}^{m}\{|\phi(b_j,y(b_j))-\phi(a_j,y(b_j))|+|\phi(a_j,y(b_j))-\phi(a_j,y(a_j))|\}\\
&\quad\leqq\sum_{j=1}^{m}\sup_{\xi\in K}|\phi(b_j,\xi)-\phi(a_j,\xi)|+l\sum_{j=1}^{m}|y(b_j)-y(a_j)|\\
&\quad<\varepsilon\quad\text{(by (6.2) and (6.3)).}
\end{aligned}$$

From elementary real analysis, we conclude that $t\to\phi(t,y(t))$ is absolutely continuous on $[0,1]$. □

Remark. As mentioned in the beginning of § 2, the choice of a function class for verification functions is a delicate matter. Basically, any class to which the value function belongs, and whose members satisfy the conclusion of Lemma 6.1 along with the appropriate monotonicity and boundary conditions, will do. It is not known whether Definition 2.2(a) could be replaced by a simpler condition; for example, an obvious candidate would be to let the sup over $\xi\in K$ be taken outside rather than inside the summation in Definition 2.1.

Proof of equation (2.2). We next prove the maximality property of V described by (2.2). Suppose $\phi:[0,1]\times\Re^n\to\Re^1$ is an arbitrary but fixed verification function. We show that if $(t,\xi)\in[0,1)\times\Re^n$, then

$$V(t,\xi)\geqq\phi(t,\xi) \tag{6.4}$$

holds, which implies the validity of (2.2) because it has already been shown that V is a verification function.

Let $(t, \xi) \in [0, 1) \times \Re^n$ and $\varepsilon > 0$. There exists $x(\cdot) \in S(t, 1, \xi)$ so that $V(t, \xi) + \varepsilon \geqq f(x(1))$. We have

$$\begin{aligned} \varepsilon + V(t, \xi) - \phi(t, \xi) &\geqq f(x(1)) - \phi(t, \xi) \\ &= \phi(1, x(1)) - \phi(t, \xi) \quad \text{(by (c))} \\ &= \int_t^1 \frac{d}{ds} \phi(s, x(s))\, ds \quad \text{(by Lemma 6.1).} \end{aligned}$$

Now suppose $s \in [t, 1]$ is such that the derivatives $\dot{x}(s)$ and $(d/ds)\phi(s, x(s))$ exist, $\dot{x}(s) \in F(t, x(s))$ and ϕ is a lower Dini solution of the Hamilton-Jacobi inequality. Then

$$\begin{aligned} \frac{d}{ds} \phi(s, x(s)) &= d^- \phi((s, x(s)); (1, \dot{x}(s))) \quad \text{(by Lemmas 5.1 and 5.2)} \\ &\geqq \min_{v \in F(s, x(s))} d^- \phi((s, x(s)); (1, v)) \geqq 0. \end{aligned}$$

Since points s with the above properties have full measure it follows that

$$\varepsilon + V(t, \xi) - \phi(t, \xi) \geqq 0.$$

In view of the fact that ε is arbitrary, (6.4) follows. □

Proof of Theorem 2.4. Suppose $x(\cdot)$ is an optimal trajectory to (P). Then the value function is a verification function with $V(0, x_0) = f(x(1))$. Conversely, suppose $x(\cdot) \in S(0, 1, x_0)$ and a verification function ϕ are given such that $\phi(0, x_0) = f(x(1))$. If $y(\cdot)$ is any other element of $S(0, 1, x_0)$, then we have

$$\begin{aligned} f(y(1)) - f(x(1)) &= \phi(1, y(1)) - \phi(0, x_0) \quad \text{(by (c))} \\ &= \int_0^1 \frac{d}{ds} \phi(x, y(s))\, ds \quad \text{(by Lemma 6.1)} \\ &\geqq 0 \quad \text{(as in (6.5)-(6.7)).} \end{aligned}$$

Hence $f(x(1))$ is the minimum value of f over $R(0, 1, x_0)$, and so $x(\cdot)$ is optimal. □

REFERENCES

[1] Z. ARTSTEIN, *On the calculus of closed set-valued functions*, Indiana Univ. Math. J., 24 (1974), pp. 433-441.

[2] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, Heidelberg, 1984.

[3] V. G. BOLTYANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control, 4 (1966), pp. 326-361.

[4] L. D. BERKOVITZ, *Optimal feedback controls*, SIAM J. Control Optim., 27 (1989), pp. 991-1006.

[5] T. F. BRIDGELAND, JR., *Trajectory integrals of set-valued functions*, Pacific J. Math., 33 (1970), pp. 43-68.

[6] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[7] F. H. CLARKE AND R. B. VINTER, *Local optimality conditions and Lipschitzian solutions to the Hamilton-Jacobi equation*, SIAM J. Control Optim., 21 (1983), pp. 856-870.

[8] M. G. CRANDALL, L. C. EVANS, AND P.-L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487-502.

[9] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[10] H. FRANKOWSKA, *Optimal trajectories associated to a solution of the contingent Hamilton-Jacobi equation*, Appl. Math. Optim., 19 (1989), pp. 291-311.

[11] R. L. GONZALES, *Sur l'existence d'une solution maximale de l'équation de Hamilton-Jacobi*, C.R. Acad. Sci., 282 (1976), pp. 1287-1290.
[12] H. HERMES, *Calculus of set-valued functions and control*, J. Math. Mech., 18 (1968), pp. 47-60.
[13] A. D. IOFFE, *Calculus of Dini subdifferentials of functions and contingent coderivatives of set valued maps*, Nonlinear Anal. Theory Methods Appl., 8 (1984), pp. 517-539.
[14] H. ISHII, *Hamilton-Jacobi equations with discontinuous Hamiltonians on arbitrary open sets*, Bull. Fac. Sci. Engng. Chuo Univ., 28 (1985), pp. 33-77.
[15] P.-L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.
[16] P.-L. LIONS AND B. PERTHAME, *Remarks on Hamilton-Jacobi equations with measurable time-dependent Hamiltonians*, Nonlinear Anal. Theory Methods Appl., 11 (1987), pp. 613-621.
[17] P.-L. LIONS AND P. E. SOUGANIDIS, *Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaac's equations*, SIAM J. Control Optim., 23 (1985), pp. 566-583.
[18] H. X. PHÚ, *On the optimal control of a hydroelectric power plant*, Systems Control Lett., 8 (1987), pp. 281-288.
[19] R. T. ROCKAFELLAR, *Integral functionals, normal integrands, and measurable selections*, in Nonlinear Operators and the Calculus of Variations, L. Waelbroeck, ed., Lecture Notes in Mathematics, vol. 543, Springer-Verlag, Berlin, New York, 1976, pp. 157-207.
[20] A. I. SUBBOTIN, *A generalization of the basic equation of the theory of differential games*, Soviet Math. Dokl., 22 (1980), pp. 358-362.
[21] ———, *A generalization of the main equation of differential game theory*, J. Optim. Theory Appl., 43 (1984), pp. 103-134.
[22] E. ROXIN, *On the generalized dynamical systems defined by contingent equations*, J. Differential Equations, 1 (1965), pp. 188-205.
[23] M. M. VALADIER, *Existence globale pour les équations différentielles multivoques*, C.R. Acad. Sci., 272 (1968), pp. 474-477.
[24] R. B. VINTER, *Weakest conditions for the existence of Lipschitz continuous Krotov functions in optimal control theory*, SIAM J. Control Optim., 21 (1983), pp. 215-234.
[25] ———, *New global optimality conditions in optimal control theory*, SIAM J. Control Optim., 21 (1983), pp. 235-245.
[26] ———, *New results on the relationship between dynamic programming and the maximum principle*, Math. Control Signals Systems, 1 (1988), pp. 97-105.
[27] R. B. VINTER AND R. M. LEWIS, *A necessary and sufficient condition for optimality of dynamic programming type, making no a priori assumptions on the controls*, SIAM J. Control Optim., 16 (1978), pp. 571-583.
[28] R. B. VINTER AND P. R. WOLENSKI, *Adjoint variables and the value function in optimal control theory: the measurable case*, J. Math. Anal. Appl., to appear.
[29] P. R. WOLENSKI, *The exponential formula for the reachable set of a differential inclusion*, SIAM J. Control Optim., 28 (1990), pp. 1148-1161.
[30] ———, *A uniqueness theorem for Lipschitz differential inclusions*, J. Differential Equations, to appear.
[31] J. A. YORKE, *Differential inequalities and non-Lipschitz and scalar functions*, Math. Systems Theory, 4 (1970), pp. 140-153.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1420-1431, November 1990

008

APPROXIMATION OF THE ZAKAI EQUATION BY THE SPLITTING UP METHOD*

A. BENSOUSSAN†, R. GLOWINSKI‡, AND A. RASCANU§

Abstract. The objective of this article is to apply an operator splitting method to the time integration of the Zakaï equation. Using this approach the numerical integration can be decomposed into a stochastic step and a deterministic one, both of them much simpler to handle than the original problem. A strong convergence theorem is given, in the spirit of existing results for deterministic problems.

Key words. nonlinear filtering, fractional step methods, Zakaï equation

AMS(MOS) subject classifications. 35K99, 60H15, 65M10

Introduction. We consider in this article an approximation technique for the Zakaï equation of nonlinear filtering. For such a filtering problem, the state and measurement processes are of the form

$$dX_t = g(X_t, t)\,dt + \sigma(X_t, t)\,dV_t,$$
$$dY_t = h(X_t, t)\,dt + dW_t,$$

where X_t denotes the system state at time t, Y_t denotes the measured output of the system at time t, and $\{V_t\colon t \geqq 0\}$, $\{W_t\colon t \geqq 0\}$ are independent Brownian motions. The goal of nonlinear filtering is to determine the conditional distribution of the state at time t given the measurements up through time t. The Zakaï equation, given by

$$dy + A^*(t)y\,dt = B(t)y \cdot dW_t, \tag{1}$$

is an evolution equation for the unnormalized conditional density of the state, given the measurements. The operators A and B are defined by

$$A(t)\varphi = \sum_{i,j} a_{ij}\frac{\partial^2\varphi}{\partial x_i \partial x_j} - \sum_i g_i \frac{\partial \varphi}{\partial x_i}$$

and

$$(B(t)\varphi)(x) = \varphi(x) \cdot h(x, t),$$

with A^* denoting the formal adjoint of A, and $a_{ij} = (\frac{1}{2}\sigma\sigma^*)_{ij}$.

We apply the idea of splitting up, considering $A(t)y\,dt - B(t)y \cdot dw$ as the sum of two operators.

Hence we write a sequence of problems of the form

$$d\varphi + A^*(t)\varphi\,dt = 0$$
$$d\psi = B(t)\psi \cdot dw^*$$

* Received by the editors February 16, 1989; accepted for publication (in revised form) October 17, 1989.

† Centre de la Recherche de Mathematique de la Décision, Universite de Paris-Dauphine, 75775 Paris, Cedex 16 and the Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau, Rocquencourt, B.P. 105, 78150 le Chesnay, France. This work was supported by the United States Army Research Office under contract DAAL03-K-0138 and by National Science Foundation grant INT-8612680 of the France-United States cooperative program.

‡ Department of Mathematics, University of Houston, Houston, Texas 77204-3476 and the Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau, Rocquencourt, B.P. 105, 78150 le Chesnay, France. This work was supported by the United States Army Research Office under contract DAAL03-K-0138 and by National Science Foundation grant INT-8612680 of the France-United States cooperative program.

§ University of Iasi, Iasi, Romania.

which are considerably simpler than (1). Indeed the φ equation is deterministic and the ψ equation has a closed-form solution.

The technique of splitting up for deterministic partial differential equations has been used extensively by many authors. It might also be noted here that this technique is very much like the Trotter product formula from semigroup theory. We refer here to the work of Temam [8], which is used as the background for our developments, and also to Glowinski [3] and Marchuk [6] for applications in mathematical physics.

We refer to Legland [4] and Elliott and Glowinski [2] for semidiscretization schemes of the Zakaï equation, which, although not related to the splitting up method, bear some analogies with those developed in this article. Legland's newest work, which contains convergence arguments of a probabilistic nature for a fully discrete approximation of the Zakaï equation, also bears some similarities to the splitting-up scheme.

1. Setting of the problem.

1.1. Notation—assumptions. We make the following assumptions on functions of the state and measurement processes:

$$g \in L^\infty(R^n \times (0,\infty); R^n),\ \sigma \in L^\infty(R^n \times (0,\infty); \mathcal{L}(R^n, R^n)), \tag{1.1}$$

with g and σ Lipschitz in x, uniformly in t, and

$$h \in L^\infty(R^n \times (0,\infty); R^m). \tag{1.2}$$

Let Ω, $\mathcal{A}$, P be a probability space on which exists an m-dimensional standard Wiener process $w(t)$, and let

$$F^t = \sigma(w(s),\ s \leqq t).$$

Define the second-order differential operator

$$A(t)\varphi = -\sum_{i,j} a_{ij}\frac{\partial^2 \varphi}{\partial x_i \partial x_j} - \sum_i g_i \frac{\partial \varphi}{\partial x_i} \tag{1.3}$$

where we have set

$$a = \tfrac{1}{2}\sigma\sigma^* \ (a = \text{matrix } a_{ij}). \tag{1.4}$$

We assume that there exists an $\alpha > 0$ such that

$$a_{ij}(x,t)\xi_i\xi_j \geqq \alpha|\xi|^2, \quad \forall \xi \in R^m, \qquad \alpha > 0. \tag{1.5}$$

We shall also define the operator

$$(B(t)\varphi)(x) \equiv \varphi(x)h(x,t). \tag{1.6}$$

Formally the Zakaï equation is written as

$$\begin{aligned} &dy + A^*(t)y\,dt = B(t)y \cdot dw \\ &y(0) = y_0. \end{aligned} \tag{1.7}$$

In the next section, we detail the function space framework necessary to analyze the system (1.7). This type of setting is used in the work of Pardoux [7] and Bensoussan [1].

1.2. Functional set up. Following the variational formulation of partial differential equations (P.D.E.) due to Lions [5], we introduce the Hilbert spaces

$$H = L^2(R^n), \qquad V = H^1(R^n)$$

and identify H with its dual. We denote by V' the dual of V.

We denote by

$$(\varphi, \psi) = \int_{R^n} \varphi\psi \, dx$$

the scalar product in H, and by

$$((\varphi, \psi)) = \int_{R^n} (\varphi\psi + D\varphi \cdot D\psi) \, dx,$$

the scalar product in V. We denote the norms on H and V by $|\cdot|$ and $\|\cdot\|$, respectively. The operator D denotes the gradient. The duality between V and V' is referred to as $\langle , \rangle$.

We now write $A(t)$ in divergence form as

$$A(t) = -\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}\right) + a_i\frac{\partial}{\partial x_i}$$

where we have set

$$a_i = \frac{\partial a_{ij}}{\partial x_j} - g_i$$

and

$$A^*(t) = -\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(a_i).$$

Note that A^* is the adjoint of A in the Hilbert space H. The operator $A(t)$ belongs to $L^\infty(0, T; \mathcal{L}(V; V'))$ and satisfies the coercivity condition

$$\exists \beta > 0, \lambda \geqq 0 \ni \langle A(t)\varphi, \varphi\rangle + \lambda|\varphi|^2 \geqq \beta\|\varphi\|^2 \quad \forall \varphi \in V, \qquad t \geqq 0. \tag{1.8}$$

This is a consequence of (1.1) and (1.5). Note that A and $A^* \in \mathcal{L}(V; V')$. Now the operator $B(t) \in L^\infty(0, T, \mathcal{L}(H; H^m))$, and we may write more clearly

$$B(t)y \cdot dw = \sum_{j=1}^{m} B^j(t)y \, dw_j,$$

where $B^j(t) \in L^\infty(0, T; \mathcal{L}(H; H))$ corresponds to

$$(B^j(t)\varphi)(x) = \varphi(x)h_j(x, t).$$

We use the notation $L_F^2(0, T; V)$ to denote the Hilbert space of processes $z(t)$ with values in V such that $E\int_0^T \|z(t)\|^2 \, dt < \infty$, and such that, for almost everywhere with respect to time, $z(t)$ is F^t measurable. Naturally, we can replace V by H or any Hilbert space.

We can state the classical result of existence and uniqueness for (1.7) (cf. Pardoux [7], cf. also Bensoussan [1]).

THEOREM 1.1. *Assume* (1.1), (1.2), (1.5). *Then, for each* $y_0 \in H$, *there exists a unique solution of* (1.7) *in the functional space*

$$y(\cdot) \in L_F^2(0, T; V) \cap L^2(\Omega, \mathcal{A}, P; C(0, T; H)).$$

The equation (1.7) can be interpreted as an Ito differential in V', since

$$y(t) = y_0 - \int_0^t A^*(s)y(s) \, ds + \sum_j \int_0^t B^j(s)y(s) \, dw_j.$$

In addition, the following Ito's calculus rule holds (equivalent of an energy equality)

$$d|y(t)|^2 + 2\langle A(t)y(t), y(t)\rangle \, dt = 2\sum_j (y, B^j y) \, dw_j + \sum_j |B^j y|^2 \, dt. \tag{1.9}$$

Note that the integrand in the stochastic integral at the right-hand side of (1.9), $(y, B^j y)$, is almost surely in $L^\infty(0, T)$ but does not belong to $L^2_F(0, T)$. This is the source of technical (although not fundamental) difficulties. To avoid them, we can rely on the following additional result.

PROPOSITION 1.1. *The process* $y(\cdot)$ *satisfies*

$$y(\cdot) \in L^\infty(0, T; L^4(\Omega, \mathcal{A}, P; H)).$$

Proof. We shall derive the a priori estimate without going into the full proof of the results. From (1.9) we deduce

$$d|y(t)|^4 = 2|y(t)|^2\left[-2\langle Ay, y\rangle\, dt + \sum_j |B^j y|^2\, dt + 2\sum_j (y, B^j y)\, dw_j\right] + 4\sum_j (y, B^j y)^2\, dt.$$

From (1.8) we have, among other facts,

$$\langle Ay, y\rangle \geqq -\lambda |y|^2;$$

hence,

$$d|y(t)|^4 \leqq \left[4\lambda |y(t)|^4 + 2|y(t)|^2 \sum_j |B^j y|^2 + 4\sum_j (y, B^j y)^2\right] dt + 4|y(t)|^2 \sum_j (y, B^j y)\, dw_j.$$

Taking the mathematical expectation yields

$$\begin{aligned}\frac{d}{dt} E|y(t)|^4 &\leqq 4\lambda E|y(t)|^4 + 2E|y(t)|^2 \sum_j |B^j y|^2 + 4E\sum_j (y, B^j y)^2 \\ &\leqq kE|y(t)|^4,\end{aligned}$$

and from Gronwall's inequality, it follows that

$$E|y(t)|^4 \leqq |y_0|^4\, e^{kt},$$

which yields the desired result. □

In the following we shall replace (1.7) by

$$dy + (A^*(t)y + \mu y)\, dt = B(t)y \cdot dw, \qquad y(0) = y_0, \tag{1.10}$$

where μ is a convenient positive constant. Since we derive (1.10) from (1.7) by the transformation $y \to ye^{-\mu t}$ it suffices to consider (1.10).

2. The splitting up approximation scheme.

2.1. The algorithm. Let N be an integer, which will tend to $+\infty$, and set

$$k = \frac{T}{N+1}.$$

We shall define two processes y_{1k}, y_{2k} depending on k. We split $[0, T]$ in steps $0, k, \cdots, (N+1)k$. Consider an interval $[rk, (r+1)k[$, $r = 0 \cdots N$, then y_{1k}, y_{2k} are defined on this interval by the relations

$$\begin{aligned} &dy_{1k} + \left(A^*(t)y_{1k} + \frac{\mu}{2} y_{1k}\right) dt = 0 \\ &dy_{2k} + \frac{\mu}{2} y_{2k}\, dt = B(t)y_{2k} \cdot dw \\ &y_{1k}(rk) = y_k^r \\ &y_{2k}(rk) = y_k^{r+1/2}\end{aligned} \tag{2.1}$$

and the sequences y_k^r, $y_k^{r+1/2}$ are defined as follows:

$$\begin{aligned} y_k^{r+1/2} &= y_{1k}((r+1)k - 0) \\ y_k^{r+1} &= y_{2k}((r+1)k - 0).\end{aligned} \tag{2.2}$$

Clearly (2.1), (2.2) define completely y_{1k}, y_{2k} in $[rk, (r+1)k[$ once y_k^r is given. As a starting point we set

$$y_k^o = y_0 \tag{2.3}$$

and (2.1), (2.2) define completely y_{1k}, y_{2k} in $[0, T[$. In (2.2) μ is a parameter which will be fixed later. The processes y_{1k}, y_{2k} are right continuous and their discontinuity points are $k, \cdots Nk$ (on $[0, T[$). Since the equation for y_{1k} is deterministic we have

$$\begin{aligned} &y_k^r, y_k^{r+1/2} \text{ are } F^{kr} \text{ measurable (with values in } H) \\ &y_{1k}(t) \text{ is } F^{kr} \text{ measurable} \quad \forall t \in [kr, (k+1)r[\\ &y_{2k}(t) \text{ is } F^t \text{ measurable} \quad \forall t. \end{aligned} \tag{2.4}$$

We can state the following existence result for (2.1).

PROPOSITION 2.1. *The system* (2.1), (2.2) *defines in a unique way* y_{1k}, y_{2k} *in* $L_F^2(0, T, V)$, $L_F^2(0, T; H)$, *respectively.*

Proof. Operating successively in each interval $[rk, (r+1)k[$, the result is clear, since for y_{1k} the deterministic theory applies and for y_{2k} we have an explicit formula for the solution. □

2.2. A priori estimates. We begin by establishing a priori estimates.

PROPOSITION 2.2. *The processes* y_{1k}, y_{2k} *satisfy*

$$E\int_0^T \|y_{1k}\|^2\, dt \leqq C, \quad E\int_0^T |y_{1k}|^2\, dt \leqq C \tag{2.5}$$

$$E|y_{1k}(t)|^4, \quad E|y_{2k}(t)|^4 \leqq C, \quad \forall t \in [0, T[, \tag{2.6}$$

where C *does not depend on* T *or* k (*for a convenient choice of* μ).

Proof. We can write the energy equalities

$$d|y_{1k}|^2 + (\mu|y_{1k}|^2 + 2\langle Ay_{1k}, y_{1k}\rangle)\, dt = 0 \tag{2.7}$$

$$d|y_{2k}|^2 + (\mu|y_{2k}|^2 - \sum_j |B^j y_{2k}|^2)\, dt = 2(y_{2k}, B^j y_{2k})\, dw_j \tag{2.8}$$

on $t \in [rk, (r+1)k[$.

We choose μ such that

$$\mu > 2\lambda, \qquad \mu > \sum_j \sup |h_j(x, t)|^2;$$

hence we deduce from (2.7), (2.8) that

$$d(\|y_{1k}(t)\|^2 + |y_{2k}(t)|^2) + \mu(\|y_{1k}\|^2 + |y_{2k}|^2)\, dt \leqq 2(y_{2k}, B^j y_{2k})dw_j, \quad \text{for each } j.$$

Integrating between $[rk, (r+1)k[$ and taking the mathematical expectation yields,

$$\begin{aligned} &E(|y_{1k}((r+1)k-0)|^2 + |y_{2k}((r+1)k-0)|^2) - E(|y_{1k}(rk)|^2 + |y_{2k}(rk)|^2) \\ &+ \mu E\int_{rk}^{(r+1)k} (\|y_{1k}\|^2 + |y_{2k}|^2)\, dt \leqq 0. \end{aligned} \tag{2.9}$$

Using (2.2) we deduce

$$E|y_k^{r+1}|^2 - E|y_k^r|^2 + \mu E\int_{rk}^{(r+1)k} (\|y_{1k}\|^2 + |y_{2k}|^2)\, dt \leqq 0. \tag{2.10}$$

Adding up the relations (2.10) for $r=0, \cdots, N$, we easily deduce

$$E\int_0^T \|y_{1k}\|^2\,dt \leqq C, \quad E\int_0^T |y_{2k}|^2\,dt \leqq C, \qquad E|y_k^r|^2 \leqq C, \tag{2.11}$$

where C does not depend on k, nor T, but only on y_0 and μ. Now using (2.7) only, integrated over $[rk, (r+1)k[$, and taking the mathematical expectation yields

$$E|y_k^{r+1/2}|^2 \leqq E|y_k^r|^2 \tag{2.12}$$

and thus also

$$E|y_k^{r+1/2}|^2 \leqq C. \tag{2.13}$$

Similarly

$$E|y_{1k}(t)|^2 \leqq E|y_k^r|^2 \leqq C \qquad \text{for } t \in [rk, (r+1)k[$$
$$E|y_{2k}(t)|^2 \leqq E|y_k^{r+1/2}|^2 \leqq C \qquad \text{for } t \in [rk, (r+1)k[.$$

Therefore we have proven that

$$E|y_{1k}(t)|^2, E|y_{2k}(t)|^2 \leqq C, \quad \forall t \in [0, T[. \tag{2.14}$$

We proceed now from (2.7), (2.8) to derive:

$$\begin{aligned} &d|y_{1k}(t)|^4 + 2|y_{1k}(t)|^2(\mu|y_{1k}|^2 + 2\langle Ay_{1k}, y_{1k}\rangle)\,dt = 0 \\ &d|y_{2k}(t)|^4 + \left[2|y_{2k}(t)|^2(\mu|y_{2k}|^2 - \sum_j |B^j y_{2k}|^2) - 4\sum_j (y_{2k}, B^j y_{2k})^2\right] dt \\ &\quad = 4|y_{2k}|^2(y_{2k}, B^j y_{2k})\,dw_j, \end{aligned} \tag{2.15}$$

and if μ is slightly larger than before, in particular satisfying

$$\mu \geqq 3\sum_j \sup |h_j|^2.$$

We derive from (2.15) that

$$E|y_{1k}((r+1)k-0)|^4 + E|y_{2k}((r+1)k-0)|^4 \leqq E|y_{1k}(rk)|^4 + E|y_{2k}(rk)|^4.$$

Hence also

$$E|y_k^{r+1}|^4 \leqq E|y_k^r|^4.$$

Therefore,

$$E|y_k^r|^4 \leqq C.$$

From this and (2.15) we easily deduce

$$E|y_{1k}(t)|^4 \leqq C, \qquad E|y_{2k}(t)|^4 \leqq C$$

and the proof of (2.5), (2.6) has been completed.

3. Convergence.

3.1. Statement of the main result. Our main result is the following theorem.

THEOREM 3.1. *Assume* (1.1), (1.2), (1.5). *Then we have*:

$$y_{1k}, y_{2k} \to y \quad \textit{in} \quad L_F^2(0, T; V) \quad \textit{and} \quad L_F^2(0, T; H), \ \textit{respectively}; \tag{3.1}$$

$$\begin{aligned} &y_{1k}(t), y_{2k}(t) \to y(t) \quad \textit{in} \quad L^2(\Omega, \mathcal{A}, P; H) \quad \forall t \in [0, T[. \\ &y_{1k}(T-0), y_{2k}(T-0) \to y(T) \quad \textit{in} \quad L^2(\Omega, \mathcal{A}, P; H). \end{aligned} \tag{3.2}$$

3.2. Weak convergence. We can extract subsequences, still denoted y_{1k}, y_{2k} such that

$$y_{1k} \to y_1 \quad \text{in} \quad L^2_F(0, T; V) \text{ weakly},$$
$$y_{2k} \to y_2 \quad \text{in} \quad L^2_F(0, T; H) \text{ weakly},$$

and

$$y_{1k}, y_{2k} \to y_1, y_2 \quad \text{in} \quad L^\infty(0, T; L^4(\Omega, \mathcal{A}, P; H)) \text{ weak star}.$$

We first have the following lemma.

LEMMA 3.1. *The functions y_1 and y_2 are equal to a common function η.*

Proof. Consider (2.1). We can integrate the first equation backward (since it is deterministic), to obtain

$$y_k^{r+1/2} - y_{1k}(t) + \int_t^{(r+1)k} \left(\frac{\mu}{2} y_{1k} + A^* y_{1k}\right) ds = 0,$$

and integrating the second equation forward, we have also

$$y_{2k}(t) - y_k^{r+1/2} + \int_{rk}^t \frac{\mu}{2} y_{2k}\, ds = \int_{rk}^t \sum_j B^j y_{2k}\, dw_j.$$

Adding up, we get (recall $t \in [rk, (r+1)k[$)

$$(y_{2k}(t) - y_{1k}(t)) + \int_t^{(r+1)k} \left(\frac{\mu}{2} y_{1k} + A^* y_{1k}\right) ds + \int_{rk}^t \frac{\mu}{2} y_{2k}\, ds = \int_{rk}^t \sum_j B^j y_{2k}\, dw_j$$

from which we deduce:[1]

$$\begin{aligned}
&E\|y_{2k}(t) - y_{1k}(t)\|^2_{V'} \\
&\leqq 3E\left(\int_t^{(r+1)k} \left\|\frac{\mu}{2} y_{1k} + A^* y_{1k}\right\|_{V'} ds\right)^2 \\
&\qquad + 3E\left(\int_{rk}^t \frac{\mu}{2} \|y_{2k}\|_{V'}\, ds\right)^2 + 3E\left(\left\|\int_{rk}^t \sum_j B^j y_{2k}\, dw_j\right\|_{V'}\right)^2 \\
&\leqq CkE \int_{rk}^{(r+1)k} (\|y_{1k}\|^2 + |y_{2k}|^2)\, ds + CE \int_{rk}^{(r+1)k} |y_{2k}|^2\, ds \\
&\leqq CkE \int_0^t (\|y_{1k}\|^2 + |y_{2k}|^2)\, ds + Ck^{1/2}\left(E \int_0^t |y_{2k}|^4\, ds\right)^{1/2} \\
&\leqq Ck^{1/2}.
\end{aligned}$$

Since $y_{2k} - y_{1k} \to y_2 - y_1$ in $L^2_F(0, T; V')$ weakly, we deduce from Fatou's Lemma that

$$\int_0^t E\|y_1 - y_2\|^2_{V'}\, dt = 0;$$

hence $y_1 = y_2 = \eta$.

[1] We use successively $\|X + Y + Z\|^2 \leqq 3(\|X\|^2 + \|Y\|^2 + \|Z\|^2)$ for all X, Y, Z in a Hilbert space; the fact that $A^* \in \mathcal{L}(V; V')$; Cauchy-Schwartz inequality $|\int_a^b f dx|^2 \leqq (b-a)\int_a^b |f|^2\, dx$; $\|\phi\|_{V'} \leqq K_1 |\phi|_H \leqq K_2 \|\phi\|_V$.

Naturally $\eta \in L^2_F(0, T; V) \cap L^\infty(0, T; L^4(\Omega, \mathscr{A}, P; H))$. □

Our objective now is to check that η satisfies (1.7) and thus $\eta = y$, by the uniqueness.

LEMMA 3.2. $\eta = y$.

Proof. We write from (2.1)

$$y_k^{r+1/2} - y_k^r + \int_{rk}^{(r+1)k} \left(\frac{\mu}{2} y_{1k} + A^* y_{1k}\right) ds = 0$$

$$y_k^{r+1} - y_k^{r+1/2} + \int_{rk}^{(r+1)k} \frac{\mu}{2} y_{1k}\, ds = \int_{rk}^{(r+1)k} \sum_j B^j y_{2k}\, dw_j,$$

hence adding up

$$y_k^{r+1/2} - y_k^r + \int_{rk}^{(r+1)k} \left(\frac{\mu}{2}(y_{1k} + y_{2k}) + A^* y_{1k}\right) ds = \int_{rk}^{(r+1)k} \sum_j B^j y_{2k}\, dw_j. \tag{3.3}$$

Adding up these relations for $r = 0, \cdots, q-1$, yields

$$y_k^q - y_0 + \int_0^{qk} \left(\frac{\mu}{2}(y_{1k} + y_{2k}) + A^* y_{1k}\right) ds = \int_0^{qk} \sum_j B^j y_{2k}\, dw_j. \tag{3.4}$$

Also from (2.1) we have for $t \in [rk, (r+1)k[$,

$$y_{1k}(t) - y_k^r + \int_{rk}^t \left(A^* y_{1k} + \frac{\mu}{2} y_{1k}\right) ds = 0. \tag{3.5}$$

Let t be fixed. Apply (3.4) with $q = [t/k]$ and (3.5) with $r = [t/k]$. Adding up, we obtain:

$$y_{1k}(t) - y_0 + \int_0^t \left(A^* y_{1k} + \frac{\mu}{2} y_{1k}\right) ds + \int_0^{k[t/k]} \frac{\mu}{2} y_{2k}\, ds = \int_0^{k[t/k]} \sum_j B^j y_{2k}\, dw_j. \tag{3.6}$$

Note that

$$\begin{aligned} E\left|\int_{k[t/k]}^t y_{2k}\, ds\right|^2 &\leqq \left(t - k\left[\frac{t}{k}\right]\right)\left(\int_{k[t/k]}^t E|y_{2k}|^2\, ds\right) \\ &\leqq C(k - [t/k]) \to 0 \end{aligned} \tag{3.7}$$

$$\begin{aligned} E\left|\int_{k[t/k]}^t \sum_j B^j y_{2k}\, dw_j\right|^2 &= E\int_{k[t/k]}^t \sum_j |B^j y_{2k}|^2\, ds \\ &\leqq CE\int_{k[t/k]}^t |y_{2k}|^2\, ds \\ &\leqq C\left(t - k\left[\frac{t}{k}\right]\right)^{1/2}\left(E\int_{k[t/k]}^t |y_{2k}|^4\, ds\right)^{1/2} \\ &\leqq C\left(t - k\left[\frac{t}{k}\right]\right)^{1/2}. \end{aligned} \tag{3.8}$$

Also

$$\int_0^t \left(A^* y_{1k} + \frac{\mu}{2} y_{1k} + \frac{\mu}{2} y_{2k}\right) ds \to \int_0^t (A^*\eta + \mu\eta)\, ds \quad \text{in } L^2(\Omega, \mathscr{A}, P; V') \text{ weakly.} \tag{3.9}$$

Let us check that

$$\int_0^t \sum_j B^j y_{2k}\, dw_j \to \int_0^t \sum_j B^j \eta\, dw_j \quad \text{in } L^2(\Omega, \mathscr{A}, P; H) \text{ weakly.} \tag{3.10}$$

Since

$$E\left|\int_0^t \sum_j B^j y_{2k}\, dw_j\right|^2 \leqq C,$$

we can assert that at least for subsequences $\int_0^t \sum_j B^j y_{2k}\, dw_j \to \chi$ in $L^2(\Omega, \mathscr{A}, P; H)$ weakly. To check that χ coincides with the right-hand side of (3.10), it is sufficient to prove that

$$E\left[\left(v, \int_0^t \sum_j B^j \eta\, dw_j\right)\xi\right] = E[(v, \chi)\xi] \tag{3.11}$$

for all $v \in H$, and $\xi \in L^2(\Omega, \mathscr{A}, P)$. This follows from the separability of H. Now since $\chi \in L^2(\Omega, F^t, P; H)$ (because $L^2(\Omega, F^t, P; H)$ is a closed subspace of $L^2(\Omega, \mathscr{A}, P; H)$ in which $\int_0^t \sum_j B^j y_{2k}\, dw_j$ stands), and since $\int_0^t \sum_j B^j \eta\, dw_j$ belongs to $L^2(\Omega, F^t, P; H)$, it is sufficient to check (3.11) with $\xi \in L^2(\Omega, F^t, P)$. Now a dense subspace of $L^2(\Omega, F^t, P)$ is made of linear combinations of random variables of the form

$$\theta(t) = \exp\left(\int_0^t \beta(s) \cdot dw(s) - \frac{1}{2}\int_0^t |\beta(s)|^2\, ds\right)$$

where β is in $L^\infty(0, t; R^m)$ (and deterministic). This follows from the fact that F^t is generated by $w(s)$, $s \leqq t$. Therefore, it is sufficient to check (3.11) with $\xi = \theta(t)$ for any β fixed. But then, what we have to prove is that

$$E\left[\left(v, \int_0^t \sum_j B^j y_{2k}\, dw_j\right)\theta(t)\right] \to E\left[\left(v, \int_0^t \sum_j B^j \eta\, dw_j\right)\theta(t)\right]. \tag{3.12}$$

However we can calculate both sizes of (3.12) by Ito's calculus, and (3.12) amounts to

$$E\int_0^t \sum_j \beta_j(s)(v, B^j y_{2k}(s))\, ds \to E\int_0^t \sum_j \beta_j(s)(v, \beta^j \eta(s))\, ds,$$

which immediately follows from the weak convergence of y_{2k} to η in $L^2_F(0, T; H)$.

Collecting results we can assert from (3.6) that

$$\forall t \quad y_{1k}(t) \to y_0 - \int_0^t (A^*\eta + \mu\eta)\, ds + \int_0^t \sum_j B^j \eta\, dw_j$$

in $L^2(\Omega, \mathscr{A}, P; H)$ weakly. Since $y_{1k}(t)$ is bounded in $L^\infty(\Omega, \mathscr{A}, P; H)$ and $y_{1k}(\cdot)$ converges weakly to $\eta(\cdot)$ in $L^2_F(0, T; H)$, we necessarily have

$$\eta(t) = y_0 - \int_0^t (A^*\eta + \mu\eta)\, ds + \int_0^t \sum_j B^j_\eta\, dw_j$$

and thus $\eta = y$.

From the uniqueness of the limit, we can assert that

$$y_{1k} \to y \quad \text{in } L^2_F(0, T; V) \text{ weakly}, \qquad y_{2k} \to y \quad \text{in } L^2_F(0, T; H) \text{ weakly},$$

and both sequences also converge in $L^\infty(0, T; L^4(\Omega, \mathscr{A}, P; H))$ weak star. □

3.3. Strong convergence. Consider (2.7), (2.8) which yield, integrating between rk and $(r+1)k$ and taking the mathematical expectation,

$$E|y_k^{r+1/2}|^2 - E|y_k^r|^2 + E\int_{rk}^{(r+1)k} (\mu|y_{1k}|^2 + 2\langle Ay_{1k}, y_{1k}\rangle)\, ds = 0$$

$$E|y_k^{r+1}|^2 - E|y_k^{r+1/2}|^2 + E\int_{rk}^{(r+1)k} \left(\mu|y_{2k}|^2 - \sum_j |B^j y_{2k}|^2\right) ds = 0.$$

Adding up we get

$$E|y_k^{r+1/2}|^2 - E|y_k^r|^2 + E\int_{rk}^{(r+1)k} \left(\mu|y_{1k}|^2 + \mu|y_{2k}|^2 - \sum_j |B^j y_{2k}|^2 + 2\langle Ay_{1k}, y_{1k}\rangle\right) ds = 0. \tag{3.13}$$

Adding up these relations for $r = 0, \cdots, q-1$, yields

$$E|y_k^q|^2 - |y_0|^2 + E\int_0^{qk} \left(\mu|y_{1k}|^2 + \mu|y_{2k}|^2 - \sum_j |B^j y_{2k}|^2 + 2\langle Ay_{1k}, y_{1k}\rangle\right) ds = 0. \tag{3.14}$$

Now from (2.7) we have for $t \in [rk, (r+1)k[$,

$$E|y_{1k}(t)|^2 - E|y_k^r|^2 + E\int_{rk}^{t} (\mu|y_{1k}|^2 + 2\langle Ay_{1k}, y_{1k}\rangle)\, ds = 0. \tag{3.15}$$

Let t be fixed. Apply (3.14) with $q = [t/k]$ and (3.15) with $r = [t/k]$. Adding up, we obtain

$$E|y_{1k}(t)|^2 - |y_0|^2 + E\int_0^t (\mu|y_{1k}|^2 + 2\langle Ay_{1k}, y_{1k}\rangle)\, ds + E\int_0^{[t/k]k} (\mu|y_{1k}|^2 - \sum_j |B^j y_{1k}|^2)\, ds = 0. \tag{3.16}$$

Now consider the expression[2]

$$\begin{aligned} X_k(t) &= E|y(t) - y_{1k}(t)|^2 + 2E\int_0^t \langle A(y - y_{1k}), y - y_{1k}\rangle\, ds \\ &\quad + E\int_0^t \mu|y - y_{1k}|^2\, ds + E\int_0^{[t/k]k} \left(\mu|y - y_{2k}|^2 - \sum_j |B^j(y - y_{2k})|^2\right) ds \\ &= X_k^1(t) + X_k^2(t) + X_k^3(t) \end{aligned}$$

with

$$\begin{aligned} X_k^1(t) &= E|y(t)|^2 + 2E\int_0^t \langle Ay, y\rangle\, ds \\ &\quad + E\int_0^t \mu|y|^2\, ds + E\int_0^{[t/k]k} \left(\mu|y|^2 - \sum_j |B^j y|^2\right) ds \\ &\to E|y(t)|^2 + 2E\int_0^t \langle Ay, y\rangle\, ds + 2\mu E\int_0^t |y|^2\, ds - E\int_0^t \sum_j |B^j y|^2\, ds \\ &= |y_0|^2; \end{aligned}$$

[2] In $X_k(t)$ the terms do not go to 0 individually, at least a priori.

$$
\begin{aligned}
X_k^2(t) = &-2E(y(t), y_{1k}(t)) - 2E\int_0^t (\langle Ay, y_{1k}\rangle \\
&+ \langle Ay_{1k}, y\rangle)\, ds - 2\mu E\int_0^t (y, y_{1k})\, ds \\
&- 2\mu E\int_0^{[t/k]k} (y, y_{2k})\, ds + 2E\int_0^{[t/k]k} \sum_j (B^j y, B^j y_{2k})\, ds \\
\to &-2E|y(t)|^2 - 4E\int_0^t \langle Ay, y\rangle\, ds \\
&- 4\mu E\int_0^t |y|^2\, ds + 2E\int_0^t \sum_j |B^j y|^2\, ds \\
= &-2|y_0|^2
\end{aligned}
$$

and from (3.16)

$$X_k^3(t) = |y_0|^2.$$

Therefore $X_k(t) \to 0$, for all t. Remark that μ was chosen so that

$$\mu > \sum_j \sup |h_j|$$

which implies

$$\mu |y - y_{2k}|^2 \geqq \sum |B^i(y - y_{2k})|^2.$$

Moreover, we have $E\int_0^t |y - y_{1k}|^2 \to 0$ which yields $y_{1k} \to y$ strongly in $L_F^2(0, T; H)$ and $E|y_{1k}(t) - y(t)|^2 \to 0$, so that $y_{1k} \to y$ in $L^2(\Omega, A, P, H)$. Next the coercivity condition $\langle A(y - y_{1k}), y - y_{1k}\rangle + \lambda |y - y_{1k}|^2 \geqq \|y - y_{1k}\|^2$ gives the $L_F^2(0, T; V)$ convergence, because the left side of the inequality is controlled by $X_k(t) \to 0$.

This implies

$$
\begin{aligned}
&y_{1k} \to y \quad \text{in } L_F^2(0, T; V) \text{ strongly}, \\
&y_{1k}(t) \to y(t) \quad \text{in } L^2(\Omega, \mathcal{A}, P; H) \quad \forall t \in [0, T[, \text{ strongly}, \\
&y_{1k}(T-0) \to y(T-0) \quad \text{in } L^2(\Omega, \mathcal{A}, P; H), \text{ strongly}.
\end{aligned} \tag{3.17}
$$

Similarly we can check that the following expression analogous to (3.16) holds:

$$
\begin{aligned}
&E|y_{2k}(t)|^2 - E|y_k^{1/2}|^2 + E\int_0^t (\mu |y_{2k}|^2 - \sum_k |B^j y_{2k}|^2)\, ds \\
&\quad + E\int_k^{([t_k]+1)k} (\mu |y_{1k}|^2 + 2\langle Ay_{1k}, y_{1k}\rangle)\, ds = 0.
\end{aligned} \tag{3.18}
$$

Using this identity, and constructing an expression similar to $X_k(t)$, which is easily guessed from the structure of (3.18), we can prove the remainder of the results ($y_{2k} \to y$ in $L_F^2(0, T, H)$ strongly).

The proof of Theorem 3.1 has been completed. □

4. Remarks.

4.1. Explicit solution. The equation for y_{2k} can be explicitly solved, namely,

$$y_{2k}(x, t) = y_k^{r+1/2}(x)\, e^{-\mu/2k} \exp \int_{rk}^t \sum_j h_j(x, s)\, dw_j(s) - \frac{1}{2}\int_{rk}^t |h(x, s)|^2\, ds$$

whereas y_{1k} is solution of the classical Fokker Plank equation. therefore we can write for y_{1k} the following equation:

$$\frac{\partial y_{1k}}{\partial t}+\left(A^*(t)y_{1k}+\frac{\mu}{2}y_{1k}\right)$$

$$(4.1)\qquad =\sum_{r=1\cdots N}\delta_{rk}(t)y_{1k}(x,\,t-0)$$

$$\cdot\left\{\exp\left[-\frac{\mu k}{2}+\int_{t-k}^{t}\sum_j h_j\,dw_j-\frac{1}{2}\int_{t-k}^{t}|h|^2\,ds\right]-1\right\},\qquad y_{1k}(0)=y_0$$

which can be considered as the approximation of the original Zakaï equation.

We can understand the right-hand side as follows. For k small and assuming h continuous in time, the term within brackets is equivalent to $-(\mu k/2)+\sum_j h_j(x,\,t)\times(w_j(t)-w_j(t-k))$, up to second-order terms. Comparing with the original Zakaï equation, it means that we have replaced the term

$$y(x,\,t)\left(-\frac{\mu}{2}\,dt+\sum_j h_j(x,\,t)\,dw_j\right)$$

with the sum of impulses

$$\sum_{r=1\cdots N}\delta_{rk}(t)y(x,\,t-0)\left(-\frac{\mu k}{2}+\sum_j h_j(x,\,t)(w_j(t)-w_j(t-k))\right).$$

4.2. Fully numerical scheme. It remains, of course, to discretize completely (4.1) both in time and space. This can be done using classical tools of numerical analysis. Numerical results will be reported elsewhere.

4.3. Extension. Our variational techniques directly inspired from the deterministic case bear two serious limitations. First, g, h must be bounded, which leaves out of the framework of the linear case. More importantly, the case when there is correlation between the system noise and the observation noise, which leads to an operator B involving the gradient of y, seems to be out of the scope of our theory.

The first limitation is purely technical, and can be overcome using Sobolev spaces with weights.

Acknowledgments. Very helpful comments and suggestions from Drs. R. J. Elliott, V. Mirelli, E. Pardoux, and from the reviewer are acknowledged.

REFERENCES

[1] A. BENSOUSSAN, *On a general class of stochastic partial differential equations*, Journal of Hydrology and Hydraulics, T. E. Unny, ed., (1987), pp. 297-303.

[2] R. J. ELLIOTT AND R. GLOWINSKI, *Approximations to solutions of the Zakai filtering equation*, Stochastic Anal. Appl., 7 (1989), pp. 145-168.

[3] R. GLOWINSKI, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.

[4] F. LEGLAND, *Estimation de paramètres dans les processus stochastiques en observation incomplète*, Thèse, Université Paris Dauphine, 1981.

[5] J. L. LIONS, *Contrôle Optimal des Systêmes gouvernês par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[6] G. I. MARCHUK, *Methods of Numerical Mathematics*, Springer-Verlag, New York, 1975.

[7] E. PARDOUX, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics, 3 (1979), pp. 127-168.

[8] R. TEMAM, *Sur la stabilité et la convergence de la méthode des pas fractionnaires*, Thèse, University of Paris, 1967.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1432-1451, November 1990

009

ON ADAPTIVE STABILIZATION OF TIME-VARYING STOCHASTIC SYSTEMS*

LEI GUO†

Abstract. The basic stability issue of time-varying stochastic systems under adaptive control is studied. A difficulty arising from treating the stochastic case as compared to the deterministic case is the lack of an a priori upper bound on the sample paths of the random noise sequence. A projected gradient algorithm with small stepsize is used, avoiding possible large deviations of the estimates. It is shown that if the unknown parameters vary slowly in some sense, then an adaptive control law can be designed so that the closed-loop system is stable. Issues of performance and robustness are also discussed.

Key words. Stochastic systems, adaptive control, random parameters, stability, short memory, projection

AMS(MOS) subject classifications. primary 93C40, secondary 93E15, 62M10

1. Introduction.

The main objective in adaptive control theory is to design controllers that perform satisfactorily for systems which possess time-varying structure. However, the primary issue is to maintain closed-loop stability.

Over the past two decades, the area of adaptive control can roughly be divided into two directions: deterministic and stochastic. In deterministic adaptive control, the system under study is normally assumed to be subjected to no noise, or at most a uniformly bounded disturbance. When the adaptive controller is designed based on deterministic methods, optimality of the performance cannot be guaranteed for time-invariant plants with a nice uniformly bounded white noise disturbance. This is due to the fact that the algorithms used have the so-called *short memory property*, i.e., the adaptation gain is not vanishing. Nevertheless, algorithms of this kind have the merit that they may stabilize a time-varying system, as has been shown recently in, e.g., Tsakalis and Ioannou (1986) and Middleton and Goodwin (1988) in the deterministic framework.

In stochastic adaptive control, noise is an essential feature of the system, and it is not necessarily bounded; a standard example is the Gaussian white noise sequence. In this case, especially for the constant parameter case, it is of interest not only to guarantee stability of the closed-loop system, but to reject the noise optimally, or at least close to optimally. This is possible, because the algorithms normally used have the so-called *long memory property*, i.e., the adaptation gain tends to zero. This guarantees that no large deviations of the estimates can occur, at least for the constant parameter case. Indeed, it has been shown that in the constant parameter case, the parameter estimates in a closed-loop adaptive system can be either nearly consistent (Becker, Kumar, and Wei (1985)) or strongly consistent (Chen and Guo (1987)). However, it is this long memory property that prevents the adaptive law from being effective for general time-varying systems. Indeed, with long memory algorithms, it has been found that it is difficult to deal with time-variations which are more complicated than, for instance, those treated in Chen and Caines (1985) and Chen and Guo (1988). For these reasons it is believed that short memory algorithms may be more effective than the long memory ones in the control of more realistically modeled time-varying systems.

* Received by the editors February 15, 1989; accepted for publication (in revised form) January 9, 1990.

† Institute of Systems Science, Academia Sinica, Beijing, 100080, People's Republic of China. This work was completed while the author was with the Department of Systems Engineering, Research School of Physical Sciences, The Australian National University, Canberra, Australia.

Let us illustrate the difference between stability studies of deterministic systems and that of stochastic systems by the following example:

$$y_{k+1} = \theta_k y_k + v_{k+1}, \qquad y_0 \neq 0, \tag{1.1}$$

$$\theta_{k+1} = \alpha\theta_k + \varepsilon_k, \qquad |\alpha| < 1. \tag{1.2}$$

Assume first that $\{v_k\}$ and $\{\varepsilon_k\}$ are deterministic sequences. It is then easy to verify the following assertion: $\{y_k\}$ is bounded for any bounded sequence $\{v_k\}$ and any sequence $\{\varepsilon_k\}$ satisfying $\sup|\varepsilon_k| \leqq \sigma$ if and only if $\sigma(1-\alpha)^{-1} < 1$. Next, let us assume that $\{v_k\}$ and $\{\varepsilon_k\}$ are independent white noise sequences. In this case, necessary conditions for the boundedness of $E\|y_k\|^2$ are discussed in, e.g., Granger and Andersen (1978) and Pourahmadi (1986). However, general sufficient conditions are hard to find even for this seemingly simple problem (see Pourahmadi (1986) for related discussions). One of the difficulties is due to the possible unboundedness of the process noise. Thus, stabilizing the first-order stochastic model (1.1)–(1.2) seems to be a nontrivial task.

By injecting an adaptive control signal to the right-hand side of model (1.1), Meyn and Caines (1987) showed that (1.1) is stabilizable if α is known, $|\sigma| < 1$, and if $\{v_k\}$ and $\{\varepsilon_k\}$ are independent Gaussian white noise sequence with known variances. The noise assumptions on $\{v_k, \varepsilon_k\}$ were subsequently relaxed in Guo and Meyn (1989) by imposing only moment conditions.

Let us now consider (1.1) again but with the unknown parameter $\{\theta_k\}$ a constant plus a first-order moving average process:

$$\theta_k = \theta + \varepsilon_k + d_1\varepsilon_{k-1}, \qquad k \geqq 0. \tag{1.3}$$

Assume that $\{v_k\}$ and $\{\varepsilon_k\}$ are independent Gaussian white noise sequences with $E|\varepsilon_k|^2 = \sigma^2 > 0$ and $d_1 > 0$. Then for second-order stability of (1.1), it is necessary that (see Tjostheim (1986, p. 60))

$$\theta^2 + [1 + (d_1)^2]\sigma^2 + 2(d_1)^2\sigma^4 < 1,$$

which implies that $|\theta| < 1$ and that σ should be suitably small. In practice, it is acceptable to assume that the noise variance σ^2 is small. However, assuming the undisturbed parameter θ to be small or less than one is generally not applaudable. Again, to make the unstable open-loop time-varying stochastic system (1.1) and (1.3) stable, the use of stochastic adaptive control techniques seems to be necessary and appealing. This problem is solved as a simple example of Theorem 1 stated later in § 3.

In this paper, we consider the basic stability issue of general time-varying stochastic systems under adaptive control. The assumptions on the random noise include two important cases: bounded sequences and Gaussian sequences. We will study two classes of SISO stochastic models, although generalizations to MIMO and some other classes are straightforward. In the first class (Model 1), the parameters are assumed to be random, and only parameters in the autoregressive part are estimated, while in the second class (Model 2) the parameters are assumed to be deterministic, and parameters in both the autoregressive and exogenous parts are estimated. The remainder of the paper is organized as follows. In § 2 we describe the stochastic models that will be studied in the paper. The main stability results are stated in § 3. Section 4 establishes some inequalities and stability results for general stochastic sequences. In § 5 we present the proofs for theorems. Further discussions on performance and robustness are given in § 6. Section 7 concludes the paper.

2. Stochastic models. In this paper we will mainly consider the following two classes of time-varying stochastic models.

Model 1 (random parameter model).

$$\begin{aligned} &y_{k+1}=a_1(k)y_k+\cdots+a_p(k)y_{k-p+1}+u_k+v_{k+1}, \qquad k\geqq 0,\\ &y_k=u_k=v_k=0 \quad \forall k<0, \end{aligned} \tag{2.1}$$

where y_k, u_k, and v_k are the scalar output, input, and random noise processes, respectively, and $a_i(k)$, $1\leqq i\leqq p$, are the unknown random time-varying parameters.

Model 2 (deterministic parameter model).

$$\begin{aligned} &y_{k+1}=a_1(k)y_k+\cdots+a_s(k)y_{k-s+1}+b_1(k)u_k+\cdots+b_t(k)u_{k-t+1}+v_{k+1}, \qquad k\geqq 0,\\ &y_k=u_k=v_k=0 \quad \forall k<0, \end{aligned} \tag{2.2}$$

where $a_i(k)$, $b_j(k)$, $1\leqq i\leqq s$, $1\leqq j\leqq t$, are the unknown deterministic time-varying parameters.

Note that both Models 1 and 2 can be rewritten in the following regression form:

$$z_{k+1}=\varphi_k^{\tau}\theta_k+v_{k+1}, \tag{2.3}$$

where for Model 1, $z_{k+1}=y_{k+1}-u_k$,

$$\varphi_k=[y_k\cdots y_{k-p+1}]^{\tau}, \qquad \theta_k=[a_1(k)\cdots a_p(k)]^{\tau}, \tag{2.4}$$

while for Model 2, $z_{k+1}=y_{k+1}$, and

$$\varphi_k=[y_k\cdots y_{k-s+1}, u_k\cdots u_{k-t+1}]^{\tau}, \qquad \theta_k=[a_1(k)\cdots a_s(k), b_1(k)\cdots b_t(k)]^{\tau}. \tag{2.5}$$

Let us now introduce the assumptions on the random noise sequence $\{v_k\}$.

Noise assumption. $\{v_k, F_k\}$ is an adapted sequence where $\{F_k\}$ is a nondecreasing family of σ-algebras, and for some integer $r\geqq 0$ and deterministic positive constants ε and M_v:

$$E\{\exp[\varepsilon\|v_{k+1}\|^2]\,|\,F_{k-r}\}\leqq\exp\{M_v\} \quad \text{a.s.} \quad \forall k\geqq 0. \tag{2.6}$$

Obviously, any sequence $\{v_k\}$ which is uniformly bounded in sample path satisfies this assumption. We note also that if $\{v_k\}$ is an r-dependent sequence (i.e., for any k, $\{v_i, i\leqq k\}$ and $\{v_{i+r}, i>k\}$ are independent), then the above assumption (2.6) reduces to

$$E\{\exp[\varepsilon|v_{k+1}|^2]\}\leqq\exp\{M_v\} \quad \text{a.s.} \quad \forall k\geqq 0. \tag{2.7}$$

Let us now give an example where the noise sequence $\{v_k\}$ is unbounded almost surely.

Example 1. Let $\{v_k\}$ be the following time-varying moving average process:

$$v_k=e_k+c_1(k)e_{k-1}+\cdots+c_r(k)e_{k-r}, \qquad k\geqq 0, \tag{2.8}$$

with deterministic coefficients $\{c_i(k)\}$ satisfying

$$\sum_{i=0}^{r}|c_i(k)|^2\leqq c<\infty \quad \forall k\geqq 0, \quad (c_0(k)=1), \tag{2.9}$$

assuming that $\{e_k\}$ is a Gaussian white noise sequence with variance $\sigma^2>0$. Then

$$\limsup_{k\to\infty}\frac{|v_k|}{(2\log k)^{1/2}}\geqq\sigma \quad \text{a.s.} \tag{2.10}$$

and the noise assumption (2.6) holds for any

$$\varepsilon<\frac{1}{2c\sigma^2}, \qquad M_v\geqq\frac{\varepsilon c\sigma^2(r+1)}{1-2\varepsilon c\sigma^2}. \tag{2.11}$$

Proof. Property (2.10) follows from the conditional Borel-Cantelli lemma and the Gaussian assumption; details of the proof are omitted (see also Chow and Teicher (1978, p. 64) for a related result). Here, we will only prove that (2.6) is true for any constants ε and M_v satisfying (2.11).

Apparently, $\{v_k\}$ is an r-dependent sequence, so we need only to verify (2.7). By elementary calculations, it is easy to verify that

$$E \exp\{\varepsilon |v_k|^2\} \leqq \{E \exp[\varepsilon c (e_1)^2]\}^{r+1} \leqq \exp\left\{\frac{\varepsilon c \sigma^2}{1-2\varepsilon c \sigma^2}(r+1)\right\}. \tag{2.12}$$

Hence by (2.11) and (2.12), we see that (2.7) is true.

We remark that in the above example, the constants ε and M_v depend only on the *upper bounds* of σ, c, and r.

3. Main results. Since the conditions imposed on the time-varying parameters of Models 1 and 2 are quite different, we will consider these two models separately.

3.1. Random parameter case. The assumptions on the parameters of Model 1 are as follows.

Parameter assumption (random case). $\{\theta_k, F_k\}$ defined in (2.4) is an adapted sequence which satisfies

$$E\{\exp[M\|\theta_{k+1}\|^2] | F_{k-m}\} \leqq \exp\{M_\theta\} \quad \text{a.s.} \quad \forall k \geqq 0, \tag{3.1}$$

$$E\{\exp[M\|w_{k+1}\|^2] | F_{k-m}\} \leqq \exp\{\delta_\theta\} \quad \text{a.s.} \quad \forall k \geqq 0, \tag{3.2}$$

where w_{k+1} is the parameter variation process:

$$w_{k+1} = \theta_{k+1} - \theta_k, \qquad k \geqq 0, \tag{3.3}$$

and where $m \geqq 0$ is an integer and M, M_θ, and $\delta_\theta < 1$ are positive deterministic constants.

We now discuss this condition. Condition (3.1) means that the random process $\{\theta_k\}$ is bounded in an average sense and not necessarily bounded in sample path. In the main theorems to follow, we will actually need that the constant M is suitably large and that δ_θ is suitably small (see Remark 3.1), which means that the parameters are slowly varying in an average sense, and again, the variation is not necessarily small in sample path. In particular, these conditions do not rule out occasional but possibly large jumps of the parameter process. Let us give a concrete example.

Example 2. Let the unknown parameter θ_k be a constant vector plus a p-dimensional moving average process:

$$\theta_k = \theta + \varepsilon_k + D_1 \varepsilon_{k-1} + \cdots + D_{m-1}\varepsilon_{k-m+1}, \qquad k \geqq 0, \tag{3.4}$$

where D_i, $1 \leqq i \leqq m-1$, are deterministic matrices, and $\{\varepsilon_k\}$ is a Gaussian white noise sequence with covariance matrix $(\sigma_\varepsilon)^2 I$. Then for any $\sigma_\varepsilon > 0$,

$$\limsup_{k\to\infty} \|\theta_k\| = \infty \quad \text{a.s.}, \qquad \limsup_{k\to\infty} \|\theta_k - \theta_{k-1}\| = \infty, \quad \text{a.s.} \tag{3.5}$$

Furthermore, the above parameter assumption holds for all small σ_ε.

Proof. We need only to verify (3.1) and (3.2) here. Note that both the process $\{\theta_k\}$ and its variation process

$$w_k = \varepsilon_k + (D_1 - 1)\varepsilon_{k-1} + \cdots + (D_{m-1} - D_{m-2})\varepsilon_{k-m+1} - D_{m-1}\varepsilon_{k-m} \tag{3.6}$$

are m-dependent sequences, so it suffices to verify (3.1) and (3.2) with conditional expectation replaced by expectation.

Similar to the proof of Example 1, we have for any constant $M>0$,

$$E\{\exp[M\|\theta_k\|^2]\}\leqq\exp\left\{2M\left[\|\theta\|^2+\frac{pmd_0(\sigma_\varepsilon)^2}{1-4Md_0(\sigma_\varepsilon)^2}\right]\right\},$$

$$E\exp\{M\|w_k\|^2\}\leqq\exp\left\{\frac{p(m+1)Md_1(\sigma_\varepsilon)^2}{1-2Md_1(\sigma_\varepsilon)^2}\right\},$$

where

$$d_0=\sum_{i=0}^{m-1}\|D_i\|^2,\quad d_1=1+\sum_{i=1}^{m}\|D_i-D_{i-1}\|^2,\quad (D_0=I,\ D_m=0).$$

Hence (3.1) and (3.2) hold. □

We now describe the estimation algorithm. Let $L>0$ and $d>0$ be two constants (which will be specified later). We define D as the following bounded domain:

$$D=\{x=(x_1,\cdots,x_p)\in R^p\colon |x_i|\leqq L,\ 1\leqq i\leqq p\} \tag{3.7}$$

and $\pi_D\{x\}$ as the nearest point from x to D (under the Euclidean norm).

The estimate for the unknown process $\{\theta_k\}$ is generated by the following projected version of the gradient algorithm:

$$\hat{\theta}_{k+1}=\pi_D\left\{\hat{\theta}_k+\frac{\varphi_k}{d+\|\varphi_k\|^2}(y_{k+1}-u_k-\varphi_k^\tau\hat{\theta}_k)\right\} \tag{3.8}$$

with arbitrary initial condition $\hat{\theta}_0\in D$, where φ_k is defined as in (2.4).

We remark that the use of a projection in estimation algorithms is common in the literature (e.g., Ljung and Soderstrom (1983), Goodwin and Sin (1984)). However, in estimating the parameters of stochastic systems by short memory algorithms, this procedure seems to be particularly important, since otherwise large deviations of the estimates are inevitable even if the system is persistently excited (see, e.g., Guo, Moore, and Xia (1988)). We also note that due to the special form of the domain D, the calculation of the projection in (3.8) is straightforward.

The certainty equivalent minimum variance adaptive control law is

$$u_k=-\varphi_k^\tau\hat{\theta}_k. \tag{3.9}$$

Our first stability result is the following theorem.

THEOREM 1. *For the random parameter model* (2.1), *if the noise assumption* (2.6) *and the parameter assumptions* (3.1)-(3.2) *hold for suitably large M and small δ_θ, and if in the estimation algorithm* (3.7)-(3.8), *L and d are taken appropriately large, then under the adaptive control law* (3.9), *the closed-loop system is stable in the sense that*

$$\limsup_{n\to\infty}E\{|y_n|^\beta+|u_n|^\beta\}<\infty, \tag{3.10a}$$

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N}\{|y_n|^2+|u_n|^2\}<\infty\quad a.s., \tag{3.10b}$$

where $\beta>2$ is a constant depending on M, δ_θ, L, and d.

Remark 3.1. We may ask how large (small) the constant M (δ_θ) is required to be in the above theorem. In § 5, we will prove that Theorem 1 is true when M and δ_θ satisfy the following inequality:

$$M\geqq 3(m+1)2^7\beta p^{3/2}\lambda^{-p}$$

and

$$\delta_\theta < \min\left\{1, \left[\frac{\beta(m+1)}{2}(\log \lambda^{-1})\right]^2\left[\frac{f(2^7(m+1)\beta p^{3/2}L_0\lambda^{-p}M^{-1/2})}{2} + \frac{2^5\beta p\lambda^{-p}(m+1)}{M}\right]^{-2}\right\} \tag{3.11}$$

for some $\lambda \in (0, 1)$ and $\beta > 2$, where the function $f(\cdot)$ is defined as

$$f(x) = \tfrac{1}{2} + x + 4x^2[1 + \exp(8x^2)] \quad \forall x \tag{3.12}$$

and L_0 denotes

$$L_0 = \left\{\frac{4M_\theta}{M} + \lambda^p[96(m+1)\beta p]^{-1}\left|\log\left[\frac{(m+1)\beta}{2}\log(\lambda^{-1})\right]\right|\right\}^{1/2}. \tag{3.13}$$

Moreover, in practical implementations of the algorithm, it is desirable to know the values of L and d. It will also be proved in § 5 that one way to choose L and d is

$$L = L_0,$$
$$d > 16p(\varepsilon\lambda^p)^{-1}\max\{8M_v(\log\lambda^{-1})^{-1}, 4\beta(r+1)\}. \tag{3.14}$$

3.2. Deterministic parameter case. The assumptions on the parameters of Model 2 are as follows.

Parameter assumption (deterministic case). (i) There is a positive constant $b_1 > 0$, such that

$$b_1(k) \geqq b_1 \quad \forall k \geqq 0, \tag{3.15}$$

and the model (2.2) is uniformly stably invertible in the sense that there are two constants $A > 0$, $\rho \in (0, 1)$ such that

$$|u_k|^2 \leqq A\sum_{i=0}^{k+1}\rho^{k+1-i}\{|y_i|^2 + |v_i|^2\} \quad \forall k. \tag{3.16}$$

(ii) The parameter is slowly varying in the sense that

$$\|\theta_k\| \leqq M_1, \quad \|\theta_{k+1} - \theta_k\| \leqq \delta_1 \quad \forall k \geqq 0, \tag{3.17}$$

where

$$\begin{aligned} &M_1 < \infty, \qquad \delta_1 < \min\{1, \log(\lambda^{-1})[24K_1\lambda^{-1}(4(s+t)^{1/2}M_1+1)]^{-1}\}, \\ &K_1 = s\lambda^{-s+1}[1 + s(M_1/b_1)^2] + (t-1)\lambda^{-t+1}[1 + (t-1)(M_1/b_1)^2]A\lambda^2/(\lambda-\rho), \end{aligned} \tag{3.18}$$

and $\lambda \in (\rho, 1)$ is some constant.

We remark that since $\{\theta_k\}$ is bounded, the assumption (3.16) is implied by uniform asymptotic stability of the following time-varying polynomial:

$$B_k(z) = b_1(k) + b_2(k)z + \cdots + b_t(k)z^{t-1}, \tag{3.19}$$

which in the constant parameter case is the standard minimum phase condition.

Let us introduce the following bounded domain:

$$D = \{x = (x_1, \cdots, x_{s+t}) \in R^{s+t} : |x_i| \leqq L,\ 1 \leqq i \leqq s+t,\ x_{s+1} \geqq b_1\}. \tag{3.20}$$

The estimation algorithm is also a projected gradient one:

$$\hat{\theta}_{k+1} = \pi_D\left\{\hat{\theta}_k + \frac{\varphi_k}{d + \|\varphi_k\|^2}(y_{k+1} - \varphi_k^\tau\hat{\theta}_k)\right\}, \tag{3.21}$$

where the initial condition $\hat{\theta}_0 \in D$, and φ_k is defined as in (2.5).

The certainty equivalent minimum variance adaptive control u_k at any time k is solved from the following simple equation:

$$\varphi_k^\tau \hat{\theta}_k = 0. \tag{3.22}$$

Similar to Theorem 1, we have the following result.

THEOREM 2. *For the deterministic parameter model* (2.2), *suppose that the noise assumption* (2.6) *and the parameter assumptions* (3.15)-(3.18) *hold, and that in the estimation algorithm* (3.20)-(3.21), *L is taken as* M_1 *appearing in* (3.17) *and*

$$d > 36K_1(\lambda\varepsilon_0)^{-1} \max\{\beta(r+1), 2M_v[\log(\lambda^{-1})]^{-1}\} \tag{3.23}$$

for some $\beta > 2$. *Then under the adaptive control law* (3.22), *the closed-loop system is stable in the sense that*

$$\limsup_{n\to\infty} E\{|y_n|^\beta + |u_n|^\beta\} < \infty, \tag{3.24a}$$

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} \{|y_n|^2 + |u_n|^2\} < \infty \quad a.s. \tag{3.24b}$$

We remark that a precise upper bound for the left-hand side of (3.24a), (3.24b) may be found in the proof—see § 5.

4. General lemmas. For the proof of theorems, we need some inequalities and stability results for stochastic sequences, which we will present in this section.

LEMMA 4.1. (i) (Bellman-Gronwall inequality). *Let* $\{x_k\}$, $\{f_k\}$, *and* $\{h_k\}$ *be three nonnegative sequences, and*

$$x_k \leqq f_k + \sum_{i=0}^{k-1} h_i x_i, \qquad k \geqq 0;$$

then

$$x_k \leqq f_k + \sum_{i=0}^{k-1} \prod_{j=i}^{k-1} (1+h_j) f_i, \qquad k \geqq 0. \tag{4.1}$$

(ii) *Let* $\{x_n, F_n\}$ *be an adapted sequence, and for some integer* $r \geqq 0$ *and some* $\alpha > 1$,

$$\sup_n E\{|x_{n+1}|^\alpha \mid F_{n-r}\} < \infty \quad a.s.; \tag{4.2}$$

then

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} |x_n| < \infty \quad a.s.$$

Proof. The first result is well known and can be easily proved by induction (see, e.g., Desoer and Vidyasagar (1975, p. 254)). As for the second result, we first note that for any fixed k, $0 \leqq k \leqq r$, the sequence

$$M_n = |x_{k+n(r+1)}| - E\{|x_{k+n(r+1)}| \mid F_{k+(n-1)(r+1)}\}$$

is a martingale difference sequence with respect to $\{F_{k+n(r+1)}\}$. Hence by (4.2) and Chow's martingale convergence theorem (see Stout (1974, p. 137)), we know that

$$\frac{1}{N} \sum_{n=0}^{N} M_n \to 0 \quad \text{a.s.} \quad \text{as } N \to \infty.$$

Consequently, by (4.2) again,

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N}|x_{k+n(r+1)}|<\infty \quad \text{a.s.} \quad \forall k\in[0,r].$$

Finally, the desired result follows by observing

$$\frac{1}{N}\sum_{n=1}^{N}|x_n|\leqq\frac{1}{N}\sum_{k=0}^{r}\sum_{n=0}^{[(N+1)/(r+1)]}|x_{k+n(r+1)}|,$$

where $[(N+1)/(r+1)]$ is the integer part of $(N+1)/(r+1)$. □

We also need the following lemma.

LEMMA 4.2. *Let* $\{x_n, F_n\}$, $\{f_n, F_n\}$, *and* $\{g_n, F_n\}$ *be three adapted nonnegative sequences satisfying*

$$x_{n+1}\leqq f_{n+1}x_n+g_{n+1} \quad \forall n\geqq 0. \tag{4.3}$$

Assume that for some constants $\varepsilon_\alpha<1$, $\alpha>1$, *and* $C<\infty$,

$$\sup_n E\{(f_{n+1})^\alpha\,|\,F_n\}\leqq\varepsilon_\alpha \quad \textit{a.s.} \quad \sup_n E[(g_{n+1})^\alpha\,|\,F_n]\leqq C. \tag{4.4}$$

Then

$$\sum_{n=0}^{N}x_n=O(N), \quad \textit{a.s.} \quad \textit{as } N\to\infty. \tag{4.5}$$

Proof. Applying the Minkowski inequality to (4.3) and noting (4.4), we see that

$$\begin{aligned}\{E(x_{n+1})^\alpha\}^{1/\alpha}&\leqq\{E(f_{n+1}x_n)^\alpha\}^{1/\alpha}+\{E(g_{n+1})^\alpha\}^{1/\alpha}\\&=\{E[E[(f_{n+1})^\alpha\,|\,F_n](x_n)^\alpha]\}^{1/\alpha}+\{E(g_{n+1})^\alpha\}^{1/\alpha}\\&\leqq(\varepsilon_\alpha)^{1/\alpha}\{E(x_n)^\alpha\}^{1/\alpha}+\sup_n\{E(g_{n+1})^\alpha\}^{1/\alpha},\end{aligned}$$

from this and the fact that $(\varepsilon_\alpha)^{1/\alpha}<1$, it is easy to conclude that

$$\sup_n E(x_n)^\alpha<\infty. \tag{4.6}$$

Let us denote $M_n=x_n-E[x_n\,|\,F_{n-1}]$; then by (4.6) and the martingale stability results (Stout (1974, p. 137)), it is evident that

$$\sum_{n=0}^{N}M_n=o(N) \quad \text{a.s.}$$

Thus by (4.4) and the recursion (4.3) we have (where ε_1 is defined as $(\varepsilon_\alpha)^{1/\alpha}$),

$$\begin{aligned}\sum_{n=0}^{N}x_{n+1}&=\sum_{n=0}^{N}E[x_{n+1}\,|\,F_n]+\sum_{n=0}^{N}M_{n+1}\\&\leqq\sum_{n=0}^{N}E[f_{n+1}\,|\,F_n]x_n+\sum_{n=0}^{N}E[g_{n+1}\,|\,F_n]+o(N)\\&\leqq\varepsilon_1\sum_{n=0}^{N}x_n+O(N)\\&\leqq\varepsilon_1\sum_{n=0}^{N}x_{n+1}+\varepsilon_1x_0+O(N),\end{aligned}$$

consequently the assertion (4.5) holds since $\varepsilon_1<1$. □

LEMMA 4.3. *Let $\{f_n\}$ be a sequence of nonnegative random variables defined by*

$$f_n = \sum_{i=0}^{n-1} \lambda^{n-i} \prod_{j=i}^{n-1} x_j, \qquad f_0 = 0,$$

where $\lambda \in (0,1)$ and $\{x_k, F_k\}$ is a nonnegative adapted sequence satisfying $x_k \geqq 1$, and

$$\{E[(x_{k+1})^{\alpha(r+1)} \mid F_{k-r}]\}^{1/[\alpha(r+1)]} \leqq C \quad a.s. \quad \lambda C < 1, \tag{4.7}$$

for some integer $r \geqq 0$ and some constants $C > 0$ and $\alpha \geqq 1$, then,

$$\sup_n \{E[f_n]^\alpha\}^{1/\alpha} \leqq \lambda C^{r+2}(1-\lambda C)^{-1}. \tag{4.8}$$

Moreover, if in (4.7) *$\alpha > 1$, then as $N \to \infty$,*

$$\sum_{n=0}^{N} f_n = O(N) \quad a.s. \tag{4.9}$$

Proof. By the Holder inequality, we have

$$\begin{aligned} E\left\{\prod_{j=i}^{n-1} x_j\right\}^\alpha &\leqq E\left\{\prod_{j=i}^{i+r} \prod_{k=0}^{[(n-i)/(r+1)]} [x_{j+k(r+1)}]^\alpha\right\} \\ &\leqq \prod_{j=i}^{i+r} \left\{E \prod_{k=0}^{[(n-i)/(r+1)]} [x_{j+k(r+1)}]^{\alpha(r+1)}\right\}^{1/(r+1)}, \end{aligned} \tag{4.10}$$

where $[(n-i)/(r+1)]$ is the integer part of $(n-i)/(r+1)$.

Note that for each i and j,

$$\begin{aligned} &E \prod_{k=0}^{[(n-i)/(r+1)]} [x_{j+k(r+1)}]^{\alpha(r+1)} \\ &= E \prod_{k=0}^{[(n-i)/(r+1)]-1} [x_{j+k(r+1)}]^{\alpha(r+1)} E\{[x_{j+[(n-i)/(r+1)](r+1)}]^{\alpha(r+1)} \mid F_{j+\{[(n-i)/(r+1)]-1\}(r+1)}\} \\ &\leqq C^{\alpha(r+1)} E \prod_{k=0}^{[(n-i)/(r+1)]-1} [x_{j+k(r+1)}]^{\alpha(r+1)} \leqq \cdots \\ &\leqq C^{\alpha(r+1)\{[(n-i)/(r+1)]+1\}} \leqq C^{\alpha(n-i+r+1)}. \end{aligned}$$

Substituting this into (4.10) we see that

$$E\left\{\prod_{j=i}^{n-1} x_j\right\}^\alpha \leqq C^{\alpha(n-i+r+1)}.$$

Consequently by the definition of f_n and the Minkowski inequality,

$$\begin{aligned} \{E[f_n]^\alpha\}^{1/\alpha} &\leqq \sum_{i=0}^{n-1} \lambda^{n-i} \left\{E \prod_{j=i}^{n-1} (x_j)^\alpha\right\}^{1/\alpha} \\ &\leqq C^{(r+1)} \sum_{i=0}^{n-1} (\lambda C)^{n-i} \leqq \lambda C^{r+2}(1-\lambda C)^{-1}. \end{aligned}$$

We now prove (4.9). By the Holder inequality,

$$\begin{aligned} \sum_{n=0}^{N} f_n &= \sum_{n=0}^{N} \sum_{i=0}^{n-1} \lambda^{n-i} \prod_{j=i}^{n-1} x_j \\ &\leqq \sum_{n=0}^{N} \sum_{i=0}^{n-1} \lambda^{n-i} \prod_{j=0}^{r} \prod_{k=0}^{[(n-i)/(r+1)]} [x_{i+j+k(r+1)}] \\ &\leqq \prod_{j=0}^{r} \left\{\sum_{n=0}^{N} \sum_{i=0}^{n-1} \lambda^{n-i} \prod_{k=0}^{[(n-i)/(r+1)]} [x_{i+j+k(r+1)}]^{r+1}\right\}^{1/(r+1)}. \end{aligned} \tag{4.11}$$

Note that for each j

$$\sum_{i=0}^{n-1} \lambda^{n-i} \prod_{k=0}^{[(n-i)/(r+1)]} [x_{i+j+k(r+1)}]^{r+1}$$

$$\leqq \sum_{s=0}^{r} \sum_{i=0}^{[n/(r+1)]} \lambda^{n-s-i(r+1)} \prod_{k=0}^{[(n-s)/(r+1)]-i} [x_{s+j+(i+k)(r+1)}]^{r+1}, \qquad \left(\prod_{k=0}^{-1} = 1\right). \tag{4.12}$$

Let us denote

$$g_n = \sum_{i=0}^{[n/(r+1)]} \lambda^{n-s-i(r+1)} \prod_{k=0}^{[(n-s)/(r+1)]-i} [x_{s+j+(i+k)(r+1)}]^{r+1}, \qquad 0 \leqq s, j \leqq r. \tag{4.13}$$

Similar to the proof of Lemma 4.1(ii), we consider the following subsequence of $\{g_n\}$ for any fixed $t \in [0, r]$, $0 \leqq s, j \leqq r$:

$$g_{t+n(r+1)} = \sum_{i=0}^{[t/(r+1)]+n} \lambda^{t-s+(n-i)(r+1)} \prod_{k=0}^{[(t-s)/(r+1)]-i+n} [x_{s+j+(i+k)(r+1)}]^{r+1}$$

$$\leqq \sum_{i=0}^{n} \lambda^{t-s+(n-i)(r+1)} \prod_{k=0}^{n-i} [x_{s+j+(i+k)(r+1)}]^{r+1} \tag{4.14}$$

$$\triangleq M_n.$$

It is obvious that $\{M_n, G_n\}$ is an adapted sequence, where $G_n = F_{s+j+n(r+1)}$. Note also that

$$M_n = [\lambda x_{s+j+n(r+1)}]^{r+1} M_{n-1} + \lambda^{t-s} [x_{s+j+n(r+1)}]^{r+1}$$

and that by the assumption (4.7),

$$\sup_n E\{[\lambda x_{s+j+n(r+1)}]^{(r+1)\alpha} \mid G_{n-1}\} \leqq (\lambda C)^{\alpha(r+1)} < 1 \quad \text{a.s.}$$

Hence applying Lemma 4.2, we have

$$\sum_{n=0}^{N} M_n = O(N) \quad \text{a.s.}$$

$$\Rightarrow \sum_{n=0}^{N} g_n = O(N) \quad \text{a.s.} \quad \text{(since in (4.14) } t \in [0, r] \text{ is arbitrary)}$$

$$\Rightarrow \sum_{n=0}^{N} \sum_{i=0}^{n-1} \lambda^{n-i} \prod_{k=0}^{[(n-i)/(r+1)]} [x_{i+j+k(r+1)}]^{r+1} = O(N) \quad \text{a.s.} \quad \text{(by (4.12))}$$

$$\Rightarrow \sum_{n=0}^{N} f_n = O(N) \quad \text{a.s.} \quad \text{(by (4.11))}.$$

This completes the proof. □

LEMMA 4.4. *Let w and F be any random variable and σ-algebra, respectively. If*

$$E\{\exp(w^2) \mid F\} \leqq \exp(\delta) \quad \textit{a.s. for some } \delta > 0,$$

then for any real number $a > 0$,

$$E\{\exp(a|w|) \mid F\} \leqq \exp\{a\delta^{1/2} + [\tfrac{1}{2} + 4a^2(1 + \exp(8a^2))]\delta\} \quad \textit{a.s.}$$

We remark that the key point in the above upper bound is the dependence on δ. If $\delta < 1$, then the above result implies that $E\{\exp(a|w|) \mid F\} \leqq \exp\{f(a)\delta^{1/2}\}$, where $f(\cdot)$ is the function defined by (3.12).

Proof. We first note that by the Jensen's inequality,

$$E\{\exp(w^2)\,|\,F\} \geqq \exp\{E[w^2\,|\,F]\} \quad \text{a.s.}$$

so it follows from the assumption that

$$E[w^2\,|\,F] \leqq \delta \quad \text{a.s.}$$

Next, we will use the following fact that can be proven in exactly the same way as that for Lemma 4.1.1 of Stout (1974, p. 226): For any random variable Y, if $0 \leqq Y \leqq 1$, almost surely, then

$$E\{\exp(Y)\,|\,F\} \leqq \exp\{E[Y\,|\,F] + E[Y^2\,|\,F]\}.$$

Applying this we have

$$E\{\exp[4a|w|I(4a|w| \leqq 1)|F]\} \leqq \exp\{4a\delta^{1/2} + (4a)^2\delta\}.$$

Hence, by this inequality, the Schwarz inequality and the Markov inequality, we have (where $E^F(\cdot)$ denotes $E(\cdot\,|\,F)$, for simplicity)

$$\begin{aligned}
&E^F \exp(a|w|) \\
&\quad = E^F \exp\{a|w|[I(|w| \geqq 2a) + I(|w| < 2a)]\} \\
&\quad \leqq E^F \exp\left(\frac{w^2}{2}\right) \exp\{a|w| I(|w| < 2a)\} \\
&\quad \leqq \exp\left(\frac{\delta}{2}\right)\left\{E^F \exp\left\{2a|w|\left[I\left(|w| \leqq \frac{1}{4a}\right) + I\left(\frac{1}{4a} < |w| < 2a\right)\right]\right\}\right\}^{1/2} \\
&\quad \leqq \exp\left(\frac{\delta}{2}\right)\left\{E^F \exp[4a|w| I(4a|w| \leqq 1)] E^F \exp\left[4a|w| I\left(\frac{1}{4a} < |w| < 2a\right)\right]\right\}^{1/4} \\
&\quad \leqq \exp\left(\frac{\delta}{2}\right)\{\exp[4a\delta^{1/2} + (4a)^2\delta]\}^{1/4}\left\{E^F \exp\left[4a|w| I\left(\frac{1}{4a} < |w| < 2a\right)\right]\right\}^{1/4} \\
&\quad \leqq \exp\left[\frac{\delta}{2} + a\delta^{1/2} + 4a^2\delta\right]\left\{E^F I\left(|w| \leqq \frac{1}{4a}\right)\right. \\
&\qquad \left. + E^F \exp[4a|w| I(|w| < 2a)] I\left(|w| > \frac{1}{4a}\right)\right\}^{1/4} \\
&\quad \leqq \exp\left[a\delta^{1/2} + \left(\frac{1}{2} + 4a^2\right)\delta\right]\left\{1 + \exp(8a^2) P\left(|w| > \frac{1}{4a}\,|F\right)\right\}^{1/4} \\
&\quad \leqq \exp\left[a\delta^{1/2} + \left(\frac{1}{2} + 4a^2\right)\delta\right]\{1 + (4a)^2\delta \exp(8a^2)\}^{1/4} \\
&\quad \leqq \exp\left\{a\delta^{1/2} + \left[\frac{1}{2} + 4a^2 + 4a^2 \exp(8a^2)\right]\delta\right\}.
\end{aligned}$$

This completes the proof. □

5. Proof of the theorems. The proofs of Theorems 1 and 2 are divided into several lemmas.

Let us denote

$$\alpha_k = \frac{\|\varphi_k^\tau \tilde{\theta}_k\|^2}{d + \|\varphi_k\|^2}, \qquad \tilde{\theta}_k = \theta_k - \hat{\theta}_k. \tag{5.1}$$

We have Lemma 5.1.

LEMMA 5.1. *Under conditions of Theorem* 1, *the following inequality holds for any* $k \geqq 0$:

$$\alpha_k \leqq 2(\|\tilde{\theta}_k\|^2 - \|\tilde{\theta}_{k+1}\|^2) + (8/d)\|v_{k+1}\|^2 + 4\{2p^{1/2}L + \|w_{k+1}\|\}\|w_{k+1}\| + 12\{p^{1/2}L + \|\theta_k\|\}\|\theta_k\| I(\theta_k \notin D),$$

where $I(A)$ *is the indicator function of a set* A.

Proof. Let us denote $\bar{\theta}_k = \theta_k I(\theta_k \in D)$. We have

$$\|\hat{\theta}_{k+1} - \bar{\theta}_k\|^2 \leqq 4pL^2$$

and

$$\begin{aligned}\|\theta_{k+1} - \bar{\theta}_k\| &\leqq \|\theta_{k+1} - \theta_k\| + \|\theta_k I(\theta_k \notin D)\| \\ &\leqq \|w_{k+1}\| + \|\theta_k I(\theta_k \notin D)\|.\end{aligned}$$

So we have

$$\begin{aligned}\|\hat{\theta}_{k+1} - \theta_{k+1}\|^2 &= \|\hat{\theta}_{k+1} - \bar{\theta}_k + \bar{\theta}_k - \theta_{k+1}\|^2 \\ &\leqq \|\hat{\theta}_{k+1} - \bar{\theta}_k\|^2 + 4p^{1/2}L\{\|w_{k+1}\| + \|\theta_k I(\theta_k \notin D)\|\} \\ &\quad + 2\|w_{k+1}\|^2 + 2\|\theta_k I(\theta_k \notin D)\|^2 \qquad (5.2) \\ &\leqq \|\hat{\theta}_{k+1} - \bar{\theta}_k\|^2 + 2\{2p^{1/2}L + \|w_{k+1}\|\}\|w_{k+1}\| \\ &\quad + 2\{2p^{1/2}L + \|\theta_k\|\}\|\theta_k\| I(\theta_k \notin D).\end{aligned}$$

But by (3.8) and the properties of the projection we know that

$$\begin{aligned}\|\bar{\theta}_k - \hat{\theta}_{k+1}\|^2 &\leqq \left\| \bar{\theta}_k - \hat{\theta}_k - \frac{\varphi_k}{d + \|\varphi_k\|^2}[\varphi_k^\tau(\theta_k - \hat{\theta}_k) + v_{k+1}] \right\|^2 \\ &= \left\| \left(I - \frac{\varphi_k \varphi_k^\tau}{d + \|\varphi_k\|^2} \right) \tilde{\theta}_k - \left\{ \theta_k I(\theta_k \notin D) - \frac{\varphi_k v_{k+1}}{d + \|\varphi_k\|^2} \right\} \right\|^2 \\ &\leqq \|\tilde{\theta}_k\|^2 - \frac{\|\varphi_k^\tau \tilde{\theta}_k\|^2}{d + \|\varphi_k\|^2} + 2\|\theta_k\|^2 I(\theta_k \notin D) + \frac{2}{d}\|v_{k+1}\|^2 \\ &\quad + 2\|\tilde{\theta}_k\|\,\|\theta_k\| I(\theta_k \notin D) + 2\frac{\|\varphi_k^\tau \tilde{\theta}_k\|\,|v_{k+1}|}{d + \|\varphi_k\|^2}.\end{aligned}$$

Applying the following elementary inequality

$$2xy \leqq \tfrac{1}{2}x^2 + 2y^2 \quad \forall x, y$$

with

$$x = \frac{\|\varphi_k^\tau \tilde{\theta}_k\|}{(d + \|\varphi_k\|^2)^{1/2}}, \qquad y = \frac{|v_{k+1}|}{(d + \|\varphi_k\|^2)^{1/2}}$$

to the last term, we then see that

$$\begin{aligned}\|\bar{\theta}_k - \hat{\theta}_{k+1}\|^2 &\leqq \|\tilde{\theta}_k\|^2 - \frac{1}{2}\frac{\|\varphi_k^\tau \tilde{\theta}_k\|^2}{d + \|\varphi_k\|^2} + \frac{4}{d}\|v_{k+1}\|^2 + 2\{(p^{1/2}L + \|\theta_k\|)\|\theta_k\| + \|\theta_k\|^2\} I(\theta_k \notin D) \\ &\leqq \|\tilde{\theta}_k\|^2 - \frac{1}{2}\frac{\|\varphi_k^\tau \tilde{\theta}_k\|^2}{d + \|\varphi_k\|^2} + \frac{4}{d}\|v_{k+1}\|^2 + 2\{p^{1/2}L + 2\|\theta_k\|\}\|\theta_k\| I(\theta \notin D).\end{aligned}$$

Substituting this into (5.2) we have

$$\|\tilde{\theta}_{k+1}\|^2 \leqq \|\tilde{\theta}_k\|^2 - \frac{1}{2}\frac{\|\varphi_k^\tau \tilde{\theta}_k\|^2}{d+\|\varphi_k\|^2} + \frac{4}{d}\|v_{k+1}\|^2 + 2\{2p^{1/2}L + \|w_{k+1}\|\}\|w_{k+1}\| + 6\{p^{1/2}L + \|\theta_k\|\}\|\theta_k\| I(\theta_k \notin D),$$

which is tantamount to the desired result. □

In a similar way, the following lemma can also be proved.

LEMMA 5.1′. *Under the conditions of Theorem* 2,

$$\alpha_k \leqq 2(\|\tilde{\theta}_k\|^2 - \|\tilde{\theta}_{k+1}\|^2) + (6/d)\|v_{k+1}\|^2 + 2\delta_1(4(s+t)^{1/2}L + \delta_1).$$

LEMMA 5.2. *Let the closed-loop system be expressed by*

$$y_{k+1} = \varphi_k^\tau \tilde{\theta}_k + v_{k+1}.$$

Assume that there are constants $\lambda \in (0,1)$, $K_1 > 0$, $K_2 \geqq 0$ *such that*

$$\sum_{i=0}^{n} \lambda^{n-i}\|\varphi_i\|^2 \leqq \sum_{i=0}^{n} \lambda^{n-i}\{K_1(y_i)^2 + K_2(v_i)^2\} \quad \forall n \geqq 0. \tag{5.3}$$

Then for any $\beta \geqq 2$,

$$\{E\|\varphi_n\|^\beta\}^{1/\beta} \leqq K_0\left\{1 + (1-\lambda)^{-1/2}\left\{E\left[\sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}(1+2K_1\lambda^{-1}\alpha_j)^2\right]^{\beta/2}\right\}^{1/\beta}\right\}^{1/2} \quad \forall n \geqq 1,$$

$$\frac{1}{N}\sum_{n=0}^{N}\|\varphi_n\|^2 \leqq O\left(\left\{\frac{1}{N}\sum_{n=0}^{N}\sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}(1+2K_1\lambda^{-1}\alpha_j)^2\right\}^{1/2}\right) + O(1),$$

where α_j *is defined in* (5.1), *and*

$$K_0 = (1-\lambda)^{-1/2}\{[2K_1+K_2][\sigma_{2\beta}(v)]^2 + 2dK_1[2pL^2 + 2(\sigma_{2\beta}(\theta))^2]\}^{1/2},$$

$$\sigma_{2\beta}(v) = \sup_k\{E|v_k|^{2\beta}\}^{1/(2\beta)}, \qquad \sigma_{2\beta}(\theta) = \sup_k\{E|\theta_k|^{2\beta}\}^{1/(2\beta)}.$$

Proof. By the assumption it follows that

$$\begin{aligned}
\|\varphi_n\|^2 &\leqq \sum_{i=0}^{n}\lambda^{n-i}\|\varphi_i\|^2 \\
&\leqq \sum_{i=0}^{n}\lambda^{n-i}\{K_1[2\|\varphi_{i-1}^\tau\tilde{\theta}_{i-1}\|^2 + 2(v_i)^2] + K_2(v_i)^2\} \\
&= 2K_1\sum_{i=0}^{n-1}\lambda^{n-i-1}\alpha_i(\|\varphi_i\|^2 + d) + (2K_1+K_2)\sum_{i=0}^{n}\lambda^{n-i}(v_i)^2 \\
&\leqq 2K_1\sum_{i=0}^{n-1}\lambda^{n-i-1}\alpha_i\|\varphi_i\|^2 + (2K_1+K_2)\sum_{i=0}^{n}\lambda^{n-i}(v_i)^2 + 2dK_1\sum_{i=0}^{n-1}\lambda^{n-i-1}\|\tilde{\theta}_i\|^2.
\end{aligned}$$

So by Lemma 4.1(i) with $x_i = \lambda^{-i}\|\varphi_i\|^2$, it is seen that

$$\|\varphi_n\|^2 \leqq \xi_n + \sum_{i=0}^{n-1}\lambda^{n-i}\left[\prod_{j=i}^{n-1}(1+2K_1\lambda^{-1}\alpha_j)\right]\xi_i \tag{5.4}$$

where

$$\xi_i = (2K_1+K_2)\sum_{k=0}^{i}\lambda^{i-k}(v_k)^2 + 2dK_1\sum_{k=0}^{i-1}\lambda^{i-k-1}\|\tilde{\theta}_k\|^2.$$

Applying the Minkowski inequality and the Schwarz inequality to (5.4), we get

$$\begin{aligned}
\{E\|\varphi_n\|^\beta\}^{2/\beta} &\leqq \{E|\xi_n|^{\beta/2}\}^{2/\beta} + \left\{E\left[\sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}(1+2K_1\lambda^{-1}\alpha_j)\xi_i\right]^{\beta/2}\right\}^{2/\beta} \\
&\leqq \{E|\xi_n|^{\beta/2}\}^{2/\beta} + \left\{E\left[\sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}(1+2K_1\lambda^{-1}\alpha_j)^2\right]^{\beta/4}\right. \\
&\qquad \left.\cdot\sum_{i=0}^{n-1}\lambda^{n-i}\|\xi_i\|^2\right]^{\beta/4}\Bigg\}^{2/\beta} \\
&\leqq \{E|\xi_n|^{\beta/2}\}^{2/\beta} + \left\{E\left[\sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}(1+2K_1\lambda^{-1}\alpha_j)^2\right]^{\beta/2}\right. \\
&\qquad \left.\cdot E\left[\sum_{i=0}^{n-1}\lambda^{n-i}\|\xi_i\|^2\right]^{\beta/2}\right\}^{1/\beta} \\
&\leqq \{E|\xi_n|^{\beta/2}\}^{2/\beta} + \left\{E\left[\sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}(1+2K_1\lambda^{-1}\alpha_j)^2\right]^{\beta/2}\right\}^{1/\beta} \\
&\qquad \cdot\left\{\sum_{i=0}^{n-1}\lambda^{n-i}[E\|\xi_i\|^\beta]^{2/\beta}\right\}^{1/2} \\
&\leqq \left\{1+(1-\lambda)^{-1/2}\left\{E\left[\sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}(1+2K_1\lambda^{-1}\alpha_j)^2\right]^{\beta/2}\right\}^{1/\beta}\right\} \\
&\qquad \cdot\sup_{0\leqq i\leqq n}\{E[\xi_i]^\beta\}^{1/\beta}.
\end{aligned}$$

Again by the Minkowski inequality,

$$\begin{aligned}
\{E[\xi_i]^\beta\}^{1/\beta} &\leqq (2K_1+K_2)\sum_{k=0}^{i}\lambda^{i-k}\{E(v_k)^{2\beta}\}^{1/\beta} + 2dK_1\sum_{k=0}^{i-1}\lambda^{i-k-1}\{E\|\tilde{\theta}_k\|^{2\beta}\}^{1/\beta} \\
&\leqq (1-\lambda)^{-1}\{[2K_1+K_2][\sigma_{2\beta}(v)]^2 + 2dK_1[2pL^2+2(\sigma_{2\beta}(\theta))^2]\}.
\end{aligned}$$

Hence the first assertion of the lemma is true, while the second assertion can easily be proved by following the similar argument and by using (5.4), Lemma 4.1(ii), and the Schwarz inequality. □

LEMMA 5.3. *Under conditions of Theorem* 1, *the property* (5.3) *holds with* $K_1 = p\lambda^{-(p-1)}$, $K_2 = 0$. *Furthermore,*

$$\{E\|\varphi_n\|^\beta\}^{1/\beta} \leqq K_0\left\{1+(1-\lambda)^{-1/2}\prod_{k=1}^{4}\{E[I_k(n)]^{\beta/2}\}^{1/(4\beta)}\right\}^{1/2} \qquad \forall n \geqq 1, \tag{5.5}$$

$$\frac{1}{N}\sum_{n=0}^{N}\|\varphi_n\|^2 \leqq O\left(\prod_{k=1}^{4}\left\{\frac{1}{N}\sum_{n=0}^{N}I_k(n)\right\}^{1/8}\right) + O(1) \quad \text{as } N\to\infty, \tag{5.6}$$

where $I_k(n)$, $k = 1, \cdots, 4$, *are defined as*

$$I_1(n) = \sum_{i=0}^{n-1}\lambda^{n-i}\exp\{8\beta_1\|\tilde{\theta}_i\|^2\}, \qquad \beta_1 = 4K_1\lambda^{-1}, \tag{5.7}$$

$$I_2(n) = \sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}\exp\left\{\frac{32\beta_1}{d}\|v_{j+1}\|^2\right\}, \tag{5.8}$$

$$I_3(n) = \sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}\exp\{16\beta_1(2p^{1/2}L+\|w_{j+1}\|)\|w_{j+1}\|\}, \tag{5.9}$$

$$I_4(n) = \sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}\exp\{48\beta_1(p^{1/2}L+\|\theta_j\|)\|\theta_j\|I(\theta_j\notin D)\}. \tag{5.10}$$

Proof. By the definition of φ_k in (2.4) and the fact that $y_k = 0$ for $k < 0$, it is easy to see that (5.3) holds with $K_1 = p\lambda^{-(p-1)}$, $K_2 = 0$.

By Lemma 5.1 and the inequality $\log(1+x) \leqq x$, for all $x \geqq 0$, we have

$$\begin{aligned}
\prod_{j=i}^{n-1}(1+2K_1\lambda^{-1}\alpha_j)^2 &= \exp\left\{\sum_{j=i}^{n-1} 2\log(1+2K_1\lambda^{-1}\alpha_j)\right\} \\
&\leqq \exp\left\{\beta_1 \sum_{j=i}^{n-1}\alpha_j\right\}, \qquad (\beta_1 = 4K_1\lambda^{-1}), \\
&\leqq \exp\{2\beta_1\|\tilde{\theta}_i\|^2\}\exp\left\{\beta_1\sum_{j=i}^{n-1}\left[\frac{8}{d}\|v_{j+1}\|^2 + 4(2p^{1/2}L+\|w_{j+1}\|)\|w_{j+1}\|\right.\right. \\
&\qquad \left.\left. + 12(p^{1/2}L+\|\theta_j\|)\|\theta_j\| I(\theta_j \notin D)\right]\right\} \\
&\leqq \exp\{2\beta_1\|\tilde{\theta}_i\|^2\}\prod_{j=i}^{n-1}\exp\left\{\frac{8\beta_1}{d}\|v_{j+1}\|^2\right\} \\
&\quad \cdot \prod_{j=i}^{n-1}\exp\{4\beta_1(2p^{1/2}L+\|w_{j+1}\|)\|w_{j+1}\|\} \\
&\quad \cdot \prod_{j=i}^{n-1}\exp\{12\beta_1(p^{1/2}L+\|\theta_j\|)\|\theta_j\| I(\theta_j \notin D)\}. \qquad (5.11)
\end{aligned}$$

Consequently, by the Holder inequality and (5.7)-(5.10),

$$\begin{aligned}
E\left[\sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}(1+2K_1\lambda^{-1}\alpha_j)^2\right]^{\beta/2} &\leqq E\left\{\prod_{k=1}^{4} I_k(n)\right\}^{\beta/8} \\
&\leqq \left\{\prod_{k=1}^{4} E[I_k(n)]^{\beta/2}\right\}^{1/4}.
\end{aligned}$$

Substituting this into Lemma 5.2, we see that (5.5) is true, while (5.6) can be proved in a similar way by using (5.11) and the Hölder inequality. The details will not be repeated. □

LEMMA 5.3′. *Under conditions of Theorem* 2, *property* (5.3) *holds with*

$$K_1 = s\lambda^{-s+1}[1+s(M_1/b_1)^2] + (t-1)\lambda^{-t+1}[1+(t-1)(M_1/b_1)^2]A\lambda^2/(\lambda-p),$$

$$K_2 = (t-1)\lambda^{-t+1}[1+(t-1)(M_1/b_1)^2]A\lambda^2/(\lambda-p).$$

Furthermore,

$$\{E\|\varphi_n\|^\beta\}^{1/\beta} \leqq K_0\left\{1+(1-\lambda)^{-1/2}\prod_{k=1}^{3}\{E[J_k(n)]^{\beta/2}\}^{1/(3\beta)}\right\}^{1/2} \qquad \forall n \geqq 1, \tag{5.12}$$

$$\frac{1}{N}\sum_{n=0}^{N}\|\varphi_n\|^2 \leqq O\left(\prod_{k=1}^{3}\left\{\frac{1}{N}\sum_{n=0}^{N}J_k(n)\right\}^{1/6}\right) + O(1) \quad \text{as } N\to\infty, \tag{5.13}$$

where $J_k(n)$, $k = 1, 2, 3$, *are defined as*

$$J_1(n) = \sum_{i=0}^{n-1}\lambda^{n-i}\exp\{6\delta_1\beta_1(n-i)[4(s+t)^{1/2}L+\delta_1]\}, \qquad \beta_1 = 4K_1\lambda^{-1}, \tag{5.14}$$

$$J_2(n) = \sum_{i=0}^{n-1}\lambda^{n-i}\exp\{6\beta_1\|\tilde{\theta}_i\|^2\}, \tag{5.15}$$

$$J_3(n) = \sum_{i=0}^{n-1}\lambda^{n-i}\prod_{j=i}^{n-1}\exp\left\{\frac{18\beta_1}{d}\|v_{j+1}\|^2\right\}. \tag{5.16}$$

Proof. Let us write $\hat{a}_i(k)$, $\hat{b}_j(k)$, as the estimates for $a_i(k)$, $b_i(k)$ given by $\hat{\theta}_k$. Then from (3.22),

$$u_k = \frac{-1}{\hat{b}_1(k)}\{\hat{a}_1(k)y_k + \cdots + \hat{a}_s(k)y_{k-s+1} + \hat{b}_2(k)u_{k-1} + \cdots + \hat{b}_t(k)u_{k-t+1}\}.$$

Since $\hat{\theta}_k$ belongs to the domain D defined by (3.20), it follows by the Schwarz inequality that

$$|u_k|^2 \leqq \left(\frac{M_1}{b_1}\right)^2 \left\{ s \sum_{j=0}^{s-1} |y_{k-j}|^2 + (t-1) \sum_{j=1}^{t-1} |u_{k-j}|^2 \right\}.$$

Therefore, by the definition of φ_i in (2.5),

$$\begin{aligned}
\sum_{i=0}^{n} \lambda^{n-i} \|\varphi_i\|^2 &= \sum_{i=0}^{n} \lambda^{n-i} \sum_{j=0}^{s-1} |y_{i-j}|^2 + \sum_{i=0}^{n} \lambda^{n-i} \sum_{j=1}^{t-1} |u_{i-j}|^2 + \sum_{i=0}^{n} \lambda^{n-i} |u_i|^2 \\
&\leqq \left[1 + s\left(\frac{M_1}{b_1}\right)^2\right] \sum_{i=0}^{n} \lambda^{n-i} \sum_{j=0}^{s-1} |y_{i-j}|^2 \\
&\quad + \left[1 + (t-1)\left(\frac{M_1}{b_1}\right)^2\right] \sum_{i=0}^{n} \lambda^{n-i} \sum_{j=1}^{t-1} |u_{i-j}|^2 \\
&\leqq s\lambda^{-s+1}\left[1 + s\left(\frac{M_1}{b_1}\right)^2\right] \sum_{i=0}^{n} \lambda^{n-i} |y_i|^2 \\
&\quad + (t-1)\lambda^{-t+1}\left[1 + (t-1)\left(\frac{M_1}{b_1}\right)^2\right] \sum_{i=0}^{n-1} \lambda^{n-i} |u_i|^2. \qquad (5.17)
\end{aligned}$$

Note that by the assumption (3.16),

$$\begin{aligned}
\sum_{i=0}^{n-1} \lambda^{n-i} |u_i|^2 &\leqq A \sum_{i=0}^{n-1} \lambda^{n-i} \sum_{j=0}^{i+1} \rho^{i+1-j} \{|y_j|^2 + |v_j|^2\} \\
&\leqq A \sum_{j=0}^{n} \sum_{i=j-1}^{n-1} \lambda^{n-i} \rho^{i+1-j} \{|y_j|^2 + |v_j|^2\} \\
&= A\lambda \sum_{j=0}^{n} \lambda^{n-j} \sum_{i=j-1}^{n-1} \left(\frac{\rho}{\lambda}\right)^{i+1-j} \{|y_j|^2 + |v_j|^2\} \\
&\leqq \frac{A\lambda^2}{\lambda - \rho} \sum_{j=0}^{n} \lambda^{n-j} \{|y_j|^2 + |v_j|^2\}.
\end{aligned}$$

Combining this with (5.17) we see that (5.3) is true. The second assertion can easily be proved by using techniques similar to those used in Lemma 5.3. We need only to note that under the present conditions the inequality (5.11) is changed to (via Lemma 5.1′),

$$\begin{aligned}
\prod_{j=i}^{n-1} (1 + 2K_1\lambda^{-1}\alpha_j)^2 &\leqq \exp\{2\delta_1\beta_1(n-i)[4(s+t)^{1/2}L + \delta_1]\} \\
&\quad \cdot \exp\{2\beta_1\|\tilde{\theta}_i\|^2\} \prod_{j=i}^{n-1} \exp\left\{\frac{6\beta_1}{d}\|v_{j+1}\|^2\right\}.
\end{aligned}$$

The details will not be repeated here. □

We now proceed to analyze the quantities $I_k(n)$, $k = 1, \cdots, 4$, appearing in (5.7)-(5.10) by using Lemma 4.3. For this we need the following lemma.

LEMMA 5.4. *Under conditions of Theorem* 1, *the following inequalities hold* (*where* $\beta_1 = 4p\lambda^{-p}$):

$$\text{(5.18)} \quad \text{(i)} \quad \sup_{(j,\omega)} E\{\exp[(16\beta\beta_1(r+1)/d)\|v_{j+1}\|^2] \mid F_{j-r}\} < \lambda^{-\beta(r+1)/2},$$

$$\text{(5.19)} \quad \text{(ii)} \quad \sup_{(j,\omega)} E\{\exp\{8\beta\beta_1(m+1)(2p^{1/2}L + \|w_{j+1}\|)\|w_{j+1}\|\} \mid F_{j-m}\} < \lambda^{-\beta(m+1)/2},$$

$$\text{(5.20)} \quad \text{(iii)} \quad \sup_{(j,\omega)} E\{\exp[24\beta\beta_1(m+1)(p^{1/2}L + \|\theta_{j+1}\|)\|\theta_{j+1}\| I(\theta_{j+1} \notin D)] \mid F_{j-m}\} < \lambda^{-\beta(m+1)/2},$$

where j takes nonnegative integer values and ω *is the sampling point.*

Proof. (i) By the noise assumption (2.6), the choice of d in (3.14), and the Hölder inequality, we have (note that $\beta_1 = 4p\lambda^{-p}$)

$$E\{\exp[(16\beta\beta_1(r+1)/d)\|v_{j+1}\|^2] \mid F_{j-r}\} \leqq \exp\{2^6 p\lambda^{-p}\beta(r+1)M_v/(\varepsilon_0 d)\} < \lambda^{-\beta(r+1)/2}.$$

(ii) By Lemma 4.4 and the parameter assumption (3.2),

$$\begin{aligned} &E\{\exp\{2^5\beta\beta_1(m+1)p^{1/2}L\|w_{j+1}\|\} \mid F_{j-m}\} \\ &\quad = E\{\exp\{[2^7(m+1)\beta p^{3/2}\lambda^{-p}LM^{-1/2}][(M)^{1/2}\|w_{k+1}\|]\} \mid F_{k-m}\} \\ &\quad \leqq \exp\{f(2^7(m+1)\beta p^{3/2}L\lambda^{-p}M^{-1/2})\delta_\theta^{1/2}\}, \end{aligned} \tag{5.21}$$

where the function $f(\cdot)$ is defined by (3.12).

Again, by the parameter assumption (3.2) and the Hölder inequality,

$$E\{\exp\{2^4\beta\beta_1(m+1)\|w_{j+1}\|^2\} \mid F_{j-m}\} \leqq \exp\{2^6\beta p\lambda^{-p}(m+1)\delta_\theta/M\};$$

combining this with (5.21) we have via the Schwarz inequality,

$$\begin{aligned} &E\{\exp\{8\beta\beta_1(m+1)(2p^{1/2}L + \|w_{j+1}\|)\|w_{j+1}\|\} \mid F_{j-m}\} \\ &\quad \leqq \exp\{[f(2^7(m+1)\beta p^{3/2}L\lambda^{-p}M^{-1/2})/2 + 2^5\beta p\lambda^{-p}(m+1)/M]\delta_\theta^{1/2}\} \\ &\quad < \lambda^{-\beta(m+1)/2}, \end{aligned}$$

where the last inequality is derived from (3.11).

(iii) We now proceed to prove (5.20). Let us denote $b = 192\beta(m+1)p^{3/2}\lambda^{-p}$; then by the parameter assumption (3.1) and the Markov inequality,

$$\begin{aligned} &E\{\exp[48\beta\beta_1(m+1)p^{1/2}L\|\theta_{j+1}\| I(\theta_{j+1} \notin D)] \mid F_{j-m}\} \\ &\quad = E\{\exp[bL\|\theta_{j+1}\| I(\theta_{j+1} \notin D)] \mid F_{j-m}\} \\ &\quad = E\{I(\theta_{j+1} \in D) \mid F_{j-m}\} + E\{\exp[bL\|\theta_{j+1}\|] I(\theta_{j+1} \notin D) \mid F_{j-m}\} \\ &\quad \leqq 1 + \{E[\exp(2bL\|\theta_{j+1}\|) \mid F_{j-m}]\}^{1/2}\{P(\|\theta_{j+1}\| > L \mid F_{j-m})\}^{1/2} \\ &\quad = 1 + \{E[\exp(2bL\|\theta_{j+1}\|) \mid F_{j-m}]\}^{1/2}\{P[\exp(2bL\|\theta_{j+1}\|) > \exp(2bL^2) \mid F_{j-m}]\}^{1/2} \\ &\quad \leqq 1 + E[\exp(2bL\|\theta_{j+1}\|) \mid F_{j-m}]/\exp(bL^2) \\ &\quad \leqq 1 + \exp(-bL^2/2)E[\exp(2b\|\theta_{j+1}\|^2) \mid F_{j-m}] \\ &\quad \leqq 1 + \exp(-bL^2/2)\exp(2bM_\theta/M) \\ &\quad \leqq \exp\{\exp[192(m+1)\beta p^{3/2}\lambda^{-p}(2M_\theta/M - L^2/2)]\} \\ &\quad \leqq \exp\{\exp[192(m+1)\beta p\lambda^{-p}(2M_\theta/M - L^2/2)]\}, \end{aligned} \tag{5.22}$$

where for the last inequality we have used the fact that $2M_\theta/M - L^2/2 \leqq 0$, which is seen from (3.13) and the choice $L = L_0$.

Similarly, we have ($c = 192\beta(m+1)p\lambda^{-p}$),

$$
\begin{aligned}
&E\{\exp[48\beta\beta_1(m+1)\|\theta_{j+1}\|^2 I(\theta_{j+1} \notin D)] \mid F_{j-m}\} \\
&\quad = E\{\exp[c\|\theta_{j+1}\|^2 I(\theta_{j+1} \notin D)] \mid F_{j-m}\} \\
&\quad \leqq 1 + \{E\{\exp[2c\|\theta_{j+1}\|^2] \mid F_{j-m}\}\}^{1/2}\{P(\|\theta_{j+1}\|^2 > L^2 \mid F_{j-m})\}^{1/2} \\
&\quad \leqq 1 + E\{\exp[2c\|\theta_{j+1}\|^2] \mid F_{j-m}\}/\exp\{cL^2\} \\
&\quad \leqq 1 + \exp\{2cM_\theta/M - cL^2\} \\
&\quad \leqq \exp\{\exp[192\beta(m+1)p\lambda^{-p}(2M_\theta/M - L^2)]\}. \qquad (5.23)
\end{aligned}
$$

Combining (5.22) and (5.23), we obtain via the Schwarz inequality,

$$
\begin{aligned}
&E\{\exp[24(m+1)\beta\beta_1(p^{1/2}L + \|\theta_{j+1}\|)\|\theta_{j+1}\| I(\theta_{j+1} \notin D)] \mid F_{j-m}\} \\
&\quad \leqq \exp\{\exp[192(m+1)\beta p\lambda^{-p}(2M_\theta/M - L^2/2)]\} < \lambda^{-\beta(m+1)/2},
\end{aligned}
$$

where the last inequality is obtained from (3.13). This completes the proof. □

Proofs of theorems. By Lemma 4.3 (with $\alpha = \beta/2 > 1$) and Lemma 5.4, we know that the quantities $I_k(n)$, $k = 2, \cdots, 4$, defined in Lemma 5.3 satisfy

$$
\sup_n E[I_k(n)]^{\beta/2} < \infty \quad \text{and} \quad \sum_{n=0}^{N} I_k(n) = O(N) \quad \text{a.s.} \quad k = 2, 3, 4, \qquad (5.24)
$$

while for $I_1(n)$, we note that

$$
\exp\{8\beta_1\|\tilde{\theta}_i\|^2\} \leqq \exp\{16\beta_1 pL^2\}\exp\{16\beta_1\|\theta_i\|^2\}.
$$

Then by the parameter assumption (3.1) and Lemma 4.1(ii), it is easy to see that (5.24) is also true for $k = 1$. Hence by Lemma 5.3 we get

$$
\sup_n E\|\varphi_n\|^\beta < \infty \quad \sum_{n=1}^{N} \|\varphi_n\|^2 = O(N) \quad \text{a.s.} \quad \text{as } N \to \infty;
$$

combining this with (3.9) we immediately conclude that Theorem 1 holds.

In a similar way, Theorem 2 can be proved. The details will not be repeated here. □

6. Further discussions. In this section we will give some brief discussions on the issues of performance and robustness.

6.1. Performance. Since our control objective is to minimize the output process, it is natural to ask if the output "approaches zero" when both the noise and the parameter variation processes are "small." Mathematically, this needs the study of, e.g., for Model 1, the asymptotic properties of $\{y_k\}$ when $(\varepsilon)^{-1} \to 0$ (M_v fixed), and $\delta_\theta \to 0$. Note that by (3.14), d is allowed to be chosen as $d \to 0$ and $(\varepsilon d)^{-1} \to 0$.

Let us denote $\bar{\varepsilon} = (\delta_\theta, d, (\varepsilon d)^{-1})$ and parameterize the output process as $\{y_k^{\bar{\varepsilon}}\}$; then from the proof of Theorem 1, it is easy to see that

$$
\lim_{\bar{\varepsilon} \to 0} \limsup_{k \to \infty} E\|y_k^{\bar{\varepsilon}}\|^\beta = 0. \qquad (6.1)
$$

For any small but fixed $\bar{\varepsilon}$, the Markov chain ergodic theory may be applied to prove the existence of the limit $\lim_{k\to\infty} E\|y_k\|^2$ if we strengthen the assumptions. For example, if $\{v_k\}$ is a Gaussian white noise sequence, and the parameter is modeled as

in Example 2, then under the assumptions of Theorem 1, the closed-loop system equations will give rise to a Markov state process $\{\Phi_k\}$, which is, in particular, (i) weakly stochastically controllable in the sense of Meyn and Caines (1988), and (ii) bounded in probability due to Theorem 1. Thus, applying Theorem 1 (for $\beta > 2$) and the important results developed in Meyn and Caines (1988) and Meyn (1988), we know that

$$\lim_{k\to\infty} P(|y_k| > x) = \pi(|y| > x) \quad \forall x,$$

$$\lim_{k\to\infty} E|y_k|^2 = \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} |y_i|^2 = \int y^2 \, d\pi < \infty, \tag{6.2}$$

where y denotes the function $y(\cdot)$ such that $y_k = y(\Phi_k)$, and π is the invariant probability of $\{\Phi_k\}$. Detailed and further results are currently under investigation. We mention that establishing the existence of the limits in (6.2) *without* using Markov chain theory appears to be a challenging problem.

6.2. Robustness. Let us assume that in addition to the random noise $\{v_k\}$ there are unmodeled dynamics $\{\eta_k\}$ acting on the system (2.3):

$$z_{k+1} = \varphi_k^\tau \theta_k + v_{k+1} + \eta_k. \tag{6.3}$$

We assume that the unmodeled dynamics $\{\eta_k\}$ depend on the previous input-output data, and have the following time-varying upper bound (see, e.g., Ioannou and Tsakalis (1985), Chen and Guo (1988)):

$$|\eta_k| \leqq \varepsilon^* m_k, \quad m_k = \gamma m_{k-1} + \|\varphi_k\|, \quad m_0 > 0, \quad k \geqq 0, \tag{6.4}$$

where $\varepsilon^* > 0$, $\gamma \in (0, 1)$.

Similar to the normalization idea used in Ioannou and Tsakalis (1985), we replace the quantity $d + \|\varphi_k\|^2$ in (3.8) or (3.21) by $d + (m_k)^2$, and consider the following algorithm:

$$\hat{\theta}_{k+1} = \pi_D \left\{ \hat{\theta}_k + \frac{\varphi_k}{d + (m_k)^2} (z_{k+1} - \varphi_k^\tau \hat{\theta}_k) \right\}. \tag{6.5}$$

Then stability of the closed-loop system under the certainty equivalent minimum variance adaptive control law can also be established, provided that ε^* is appropriately small. The proof is essentially the same as that for Theorems 1 and 2.

7. Conclusion. In this paper, stabilizing adaptive controllers are presented for possible open-loop unstable time-varying stochastic systems described by Models 1 and 2. The closed-loop stability is proved based on an analysis of products of random variables and truncation techniques. We have seen that the use of projection in the estimation algorithm plays a crucial role in getting useful estimates in the stochastic case, especially when the noise is unbounded in sample path. Further asymptotic results are currently being explored by applying the weak convergence theory and the Markov chain ergodic theory.

Acknowledgments. The author thanks Dr. S. P. Meyn for his comments and valuable discussions on Markov chain theory. Thanks are also due to Professor P. A. Ioannou of the University of Southern California for his valuable discussions during the author's visit to the Australian National University.

REFERENCES

A. BECKER, P. R. KUMAR, AND C. Z. WEI (1985), *Adaptive control with the stochastic approximation algorithm, geometry and convergence*, IEEE Trans. Automat. Control, 30, pp. 330–338.

H. F. CHEN AND P. E. CAINES (1985), *On the adaptive control of a class of systems with random parameters and disturbances*, Automatica, 21, pp. 737–741.

H. F. CHEN AND L. GUO (1987), *Asymptotically optimal adaptive control with consistent parameter estimates*, SIAM J. Control Optim., 25, pp. 558–575.

——— (1988), *A robust stochastic adaptive controller*, IEEE Trans. Automat. Control, 33, pp. 1035–1043.

Y. S. CHOW AND H. TEICHER (1978), *Probability Theory: Independency, Interchangeability and Martingales*, Springer-Verlag, New York.

C. A. DESOER AND M. VIDYASAGAR (1975), *Feedback Systems: Input-Output Properties*, Academic Press, New York.

G. C. GOODWIN AND K. S. SIN (1984), *Adaptive Filtering, Predication and Control*, Prentice-Hall, Englewood Cliffs, NJ.

C. W. J. GRANGER AND A. P. ANDERSEN (1978), *An Introduction to Bilinear Time Series Models*, Vandenhoeck & Ruprecht, Gottingen.

L. GUO AND S. P. MEYN (1989), *Adaptive control for time-varying systems: a combination of martingale and Markov chain techniques*, Internat. J. Adaptive Control Signal Process., 3, pp. 1–14.

L. GUO, J. B. MOORE, AND L. G. XIA (1988), *Tracking randomly varying parameters—analysis of a standard algorithm*, Proc. 27th IEEE Conf. on Decision and Control, Austin, Texas, pp. 1514–1519.

P. IOANNOU AND K. TSAKALIS (1985), *Robust discrete-time adaptive control*, in Adaptive and Learning Systems; Theory and Applications, K. S. Narendra, ed., Plenum Press, New York, pp. 73–85.

L. LJUNG AND T. SODERSTROM (1983), *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA.

R. H. MIDDLETON AND G. C. GOODWIN (1988), *Adaptive control of time-varying linear systems*, IEEE Trans. Automat. Control, 33, pp. 150–157.

S. P. MEYN AND P. E. CAINES (1987), *A new approach to stochastic adaptive control*, IEEE Trans. Automat. Control, 32, pp. 220–226.

S. P. MEYN (1989), *Ergodic theorems for discrete time stochastic systems using a stochastic Lyapunov function*, SIAM J. Control Optim., 27, pp. 1409–1439.

S. P. MEYN AND P. E. CAINES (1991), *Asymptotic behavior of stochastic systems possessing Markovian realizations*, SIAM J. Control Optim., to appear.

M. POURAHMADI (1986), *On stationary of the solution of a double stochastic model*, J. Time Series Anal., 7, pp. 123–131.

W. F. STOUT (1974), *Almost Sure Convergence*, Academic Press, New York.

D. TJOSTHEIM (1986), *Some double stochastic time series models*, J. Time Series Anal., 7, pp. 51–72.

K. TSAKALIS AND P. A. IOANNOU (1986), *Adaptive control of linear time-varying plants*, in Proc. IFAC Workshop on Adaptive Systems Control and Signal Processing, Lund, Sweden.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1452-1460, November 1990

010

RELATIONSHIPS BETWEEN VARIOUS MARKOVIAN DECISION PROBLEM CLASSES*

GARY J. KOEHLER†

Abstract. In this paper Markov-type decision problems that are generalizations of discrete-time, finite state, and action stationary Markov decision problems either in allowing negative entries and/or placing less restrictions on the spectral radius are considered. Many of the results found in the literature of Markov-type decision problems were obtained by authors investigating one of several solution procedures for solving the related problem of interest. This orientation has led to a number of seemingly unrelated results. In this paper several Markov-type decision problem classes are given and it is pointed out where gaps still appear in an overall theory.

Key words. Markov decision problems, value convergence, policy iteration, linear programming, complementarity, fixed points

AMS(MOS) subject classifications. 49, 60

1. Introduction. Let $S \equiv \{1, \cdots, m\}$. For each $i \in S$, let Δ_i be a nonempty, finite, ordered set. Each $j \in \Delta_i$ indexes a one-by-m real vector p_{ij} and scalar c_{ij}. Let $\Delta \equiv X\Delta_i$. $\delta \in \Delta$ indexes a matrix P_δ and vector c_δ.

In finite-state and action Markov decision problems S is the finite state space, Δ_i is the set of (finite) actions available in state i, Δ is the set of policies, p_{ij} is the vector of transition probabilities from state i under action j, c_{ij} is the current reward in state i when action j is taken, P_δ is the transition matrix under policy δ, and c_δ is the vector of current rewards under policy δ. If the returns are discounted, then each element of p_{ij} is assumed to already contain the discount factor.

Define the affine functions $L_\delta : R^m \to R^m$ and $L : R^m \to R^m$ by

$$L_\delta(v) \equiv P_\delta v + c_\delta$$

and

$$L(v) \equiv \operatorname*{Vmax}_{\delta \in \Delta} (P_\delta v + c_\delta)$$

where Vmax means the vector (component by component) maximum. It is readily apparent that $L(\ )$ exists due to the construction and finiteness of Δ.

In finite-state and action Markov decision problems $L_\delta(v)$ is the expected (discounted) reward after one transition with policy δ and terminal rewards of v. $L(v)$ is the best expected (discounted) return after one transition with policy δ and terminal rewards of v.

Define $k \equiv \Sigma_i |\Delta_i|$ and let (A, c) be the k-by-$(m+1)$ matrix formed by $(e_i - p_{ij}, c_{ij})$ ordered by $j \in \Delta_i$, $i \in S$ where e_i is the ith unit vector. In Theorem 2 we assume that A' cannot be partitioned to render smaller independent subproblems. This is easily tested and implemented using a procedure given by Bather [1].

Let

$$F \equiv \{v:\ v = L(v)\}$$

be the set of fixed points of $L(\ )$.

* Received by the editors February 13, 1989; accepted for publication (in revised form) January 25, 1990.

† Department of Decision and Information Sciences, College of Business Administration, University of Florida, Gainesville, Florida 32611.

Consider the problems

(E) Find $f \in F$

and

(P) $v^* \equiv \underset{f \in F}{\text{Vmin}}\, f$

when the values exist. Clearly, any solution of (P) is a solution of (E). This can be seen in Cottle and Veinott [2, Thm. 2, Cor. 2, pp. 244-245].

In discounted finite-state and action Markov decision problems the solutions of $v = L_\delta(v)$ are the expected discounted rewards over an infinite horizon with stationary policy δ. v^* is the maximal such reward across all policies. The decision problems originally considered by Howard [7], Veinott [17], Jewell [8], Koehler [10], [11], and Eaves [4] fit this paradigm.

Much of the investigation of $L(\ )$ and F has been done by researchers having one of several solution procedures in mind. Their results have reflected the strengths and weaknesses of the selected method.

In this paper we review several scattered results on Markov-type decision problems and show their relationships to each other. In doing so we point out gaps which represent areas that may prove fruitful for further research.

2. Solution procedures. Let $D \equiv \{v: Av \geqq c\}$. Note that $F \subseteq D$. Let $A_\delta \equiv I - P_\delta$ for each $\delta \in \Delta$ where I is the identity matrix. Let $\rho(P)$ be the spectral radius of the square matrix P. We will draw frequently from various duality theorems [12] and the Perron-Frobenius theorem for nonnegative matrices (e.g., see [15, Thms. 1.1, 1.5]).

Below we point our four primary solution methods for solving problem (P) or (E).

DEFINITION 1 (value iteration, successive approximation). Value iteration is an algorithm defined as follows:

V1: Choose $v^0 \in R^m$.
V2: $v^{n+1} \equiv L(v^n)$.
V3: If
$$v^{n+1} \in B_\varepsilon(F)$$
then stop;
else go to V2.

Here $B_\varepsilon(T)$ is an epsilon neighborhood of set T and $\varepsilon \geqq 0$ and small enough.

DEFINITION 2 (policy iteration). Policy iteration is an algorithm defined as:

P1: Choose $\delta \in \Delta$.
P2: Solve
$$A_\delta v = c_\delta.$$
Call the solution v_δ.
P3: If
$$\delta \in \Delta(v_\delta) \equiv \arg \underset{\gamma \in \Delta}{\text{Vmax}}\, (P_\gamma v_\delta + c_\gamma)$$
then stop;
else choose $\delta \in \Delta(v_\delta)$ and go to P2.

DEFINITION 3 (linear programming). A linear programming approach for solving (P) is:

(LPP): Minimize $b'v$
subject to $Av \geqq c$

with $b > 0$. The constraints can also be written as

$$v \in D.$$

The dual is frequently solved:

(LPD): Maximize $c'x$
subject to $A'x = b, \quad x \geqq 0.$

The final method we wish to consider is the *complementarity* approach by Eaves [4]. We refer the reader to [4] for more details of the algorithm.

DEFINITION 4 (complementarity approach). Consider the system of equations

(CP): $x - Ay - dz = -c$
$x, z \geqq 0$
$\exists \delta \in \Delta$ such that $x_\delta = 0$

where x is k by 1, y is m by 1, d is k by 1, and z is a scalar variable. $d_i = 0$ whenever row i of A corresponds to the values of the first action of a state. Otherwise $d_i = 1$. If (x, y, z) is a solution to (CP) with $z = 0$, then there is a policy $\delta \in \Delta$ with $x_\delta = 0$ and $y = L(y) = L_\delta(y)$ [4, Lemma 1].

A solution to (CP)—if any—is obtained by the following algorithm. Let β_0 be an initial "basis" of (CP). (For an exact definition see Eaves [4, pp. 66-67].)

P1: determine β_0 and set $r = 0$.
P2: Assume a sequence of bases $\beta_0, \beta_1, \cdots, \beta_r$ has been constructed. If z is not in β_r, then stop. Otherwise find an "adjacent basis," β_{r+1}, other than β_{r-1}, using complementarity pivots. Increment r and go to Step P2.
Otherwise, stop.

For a given problem (Δ, A, c), none of the four methods may find a fixed point of $L(\ )$, even when F is nonempty. For example, step V3 in value iteration may never be satisfied. P2 in policy iteration may not have a solution. (LPP) may be unbounded.

3. Problem classes. Several problem classes solvable by one or more of the solution procedures defined in the preceding section are now defined.

The Markov-type decision problem defined by Veinott [17] includes traditional discounted Markov decision problems studied by others [7], [8]. Let C_1 be defined as follows.

DEFINITION 5 (class C_1). Class C_1 is the class of all problems (Δ, A, c) satisfying the following conditions:

(1) $P_\delta \geqq 0$, for all $\delta \in \Delta$.
(2) $\rho(P_\delta) < 1$, for all $\delta \in \Delta$.

THEOREM 1 (Denardo [3], Veinott [17]). *If* $(\Delta, A, c) \in C_1$, *then* F *has a unique element* v^*. *Value iteration* (*for any* $v^0 \in R^m$), *policy iteration* (*for any* $\delta \in \Delta$ *in step* P1), *and linear programming all lead to* v^*.

Problems in class C_1 are called contracting [3] or transient [17]. In finite-state and finite-action discounted Markov decision problems, each $P_\delta \equiv aQ_\delta$ where $a \in [0, 1)$ and each row of Q_δ sums to one. The spectral radius of aQ_δ satisfies $\rho(aQ_\delta) = \in < 1$. Hence these problems belong to C_1. Semi-Markov problems also fall into this category.

Koehler [11] investigated a more general problem defined as follows.

DEFINITION 6 (Class C_2). Class C_2 is the class of all problems (Δ, A, c) satisfying the following conditions:

(1) $P_\delta \geqq 0$, for all $\delta \in \Delta$.
(2) $\rho(P_\delta) < 1$, for some $\delta \in \Delta$.

(3) A_δ has a positive diagonal, for all $\delta \in \Delta$.

(4) $\rho(P_\delta) \leqq 1$, for all $\delta \in \Delta$.

(5) If $\rho(P_\delta) = 1$, then $A_\delta v = c_\delta$ has no solution.

(6) D is nonempty.

THEOREM 2 (Koehler [11]). *If* $(\Delta, A, c) \in C_2$, *then F has a unique element,* v^* *and* $A_{\delta^*} v^* = c_{\delta^*}$ *for some* δ^* *where* $\rho(P_{\delta^*}) < 1$. *Value iteration (for any* $v^0 \in R^m$*) and linear programming all lead to* v^*. *Policy iteration may fail at step* P2 *for an arbitrary starting* $\delta \in \Delta$. *However, if we choose* $\delta \in \Delta$ *at step* P1 *such that* $\rho(P_\delta) < 1$, *then policy iteration will find* v^*.

Koehler [10], [11] also investigated a more general problem defined as follows.

DEFINITION 7 (class C_3). Class C_3 is the class of all problems (Δ, A, c) satisfying the following conditions:

(1) $P_\delta \geqq 0$, for all $\delta \in \Delta$.

(2) $\rho(P_\delta) < 1$, for some $\delta \in \Delta$.

(3) D is nonempty.

THEOREM 3 (Koehler [10], [11]). *If* $(\Delta, A, c) \in C_3$, *then F is nonempty but may not have a unique element.* v^* *is F's least element.* $A_{\delta^*} v^* = c_{\delta^*}$ *for some* δ^* *where* $\rho(P_{\delta^*}) < 1$. *Value iteration may not converge or may converge to a fixed point that is not the least element of F. Policy iteration may fail at step* P2 *for an arbitrary starting* $\delta \in \Delta$. *If it does yield a fixed point, it may not be* v^*. *Linear programming will lead to* v^*. *Value iteration will lead to* v^* *when* $v^0 \leqq v^*$.

Eaves [4] considered the problem of finding a fixed point of $L(\ )$ under rather general conditions.

DEFINITION 8 (classes C_4 and C_4^+). Class C_4 is the class of all problems (Δ, A, c) satisfying the following condition:

$$\sum_{\delta \in \Delta} D_\delta A_\delta \quad \text{is nonsingular}$$

for all

$$D_\delta \geqq 0 \quad \text{and diagonal}$$

giving

$$\sum_{\delta \in \Delta} D_\delta = I.$$

C_4^+ is the subset of C_4 where $P_\delta \geqq 0$ for each $\delta \in \Delta$.

Note that A_δ must be nonsingular for each $\delta \in \Delta$.

THEOREM 4 (Eaves [4]). *If* $(\Delta, A, c) \in C_4$, *then F has a unique element* v^*. *Eaves' complementarity procedure will find* v^*.

Eaves [4] also considered a more general class of the following form.

DEFINITION 9 (classes C_5 and C_5^+). Class C_5 is the class of all problems (Δ, A, c) satisfying the following conditions:

(1) A_{δ^0} nonsingular for some $\delta^0 \in \Delta$.

(2) $\{v: A_{\delta^0} v \geqq c_{\delta^0}, A_\delta v \leqq c_\delta\}$ is bounded for each $\delta \in \Delta$.

C_5^+ is the subset of C_5 where $P_\delta \geqq 0$ for each $\delta \in \Delta$.

THEOREM 5 (Eaves [4]). *If* $(\Delta, A, c) \in C_5$, *then F is nonempty. Eaves' complementarity procedure will find an element of F. Linear programming may not find a solution. Simple examples illustrate that value iteration and policy iteration may also fail under these conditions.*

4. Relationships between problem classes. Our first result shows that $C_1 \subseteq C_2 \subseteq C_3$. We use the following property.

LEMMA 1 (Fiedler and Ptak [5]). *If $P \geqq 0$ is square and $\rho(P) < 1$, then $I - P$ has a positive diagonal and $(I-P)^{-1} \geqq 0$.*

THEOREM 6. $C_1 \subseteq C_2 \subseteq C_3$.

Proof. $(C_1 \subseteq C_2)$. Let $(\Delta, A, c) \in C_1$. Then $P_\delta \geqq 0$ and $\rho(P_\delta) < 1$ for each $\delta \in \Delta$. By Lemma 1, A_δ has a positive diagonal for each $\delta \in \Delta$. Finally, from Theorem 2 $v^* \in F \subseteq D$. So D is nonempty.

$(C_2 \subseteq C_3)$. Each property of C_3 is a property of C_2. □

It is easy to show that C_2 is not contained in C_4. C_2 permits problems where $\rho(P_\delta) = 1$ for some $\delta \in \Delta$. In such cases A_δ is singular. We noted earlier that A_δ is nonsingular for all $\delta \in \Delta$ for problems in class C_4.

THEOREM 7. $C_4 \subseteq C_5$.

Proof. Let $(\Delta, A, c) \in C_4$. Then $\Sigma D_\delta A_\delta$ is nonsingular whenever $\Sigma D_\delta = I$ and $D_\delta \geqq 0$ and diagonal. Let $\delta^0 \in \Delta$ be arbitrary and suppose that

$$\{v: A_{\delta^0} v \geqq c_{\delta^0}, A_\delta v \leqq c_\delta\}$$

is unbounded for some $\delta \in \Delta$. Thus, there is a nonzero vector r such that $A_{\delta^0} r \geqq 0$, $A_\delta r \leqq 0$. Let D_{δ^0} and D_δ be constructed so that $D_{\delta^0} + D_\delta = I$ and D_{δ^0} and D_δ are nonnegative diagonal matrices giving

$$D_{\delta^0} A_{\delta^0} r + D_\delta A_\delta r = 0.$$

Thus

$$(D_{\delta^0} A_{\delta^0} + D_\delta A_\delta) r = 0,$$

which contradicts the nonsingularity of the term $D_{\delta^0} A_{\delta^0} + D_\delta A_\delta$. Hence $(\Delta, A, c) \in C_5$. □

The same proof can be used to show that $C_4^+ \subseteq C_5^+$.

THEOREM 8. $C_1 \subseteq C_4^+ \subseteq C_4$.

Proof. The proof follows directly from results of Fiedler and Ptak [5]. □

While C_2 is not contained in C_4, it is contained in C_5.

THEOREM 9. $C_2 \subseteq C_5^+ \subseteq C_5$.

Proof. Let $(\Delta, A, c) \in C_2$ and let δ^0 be any $\delta^0 \in \Delta$ giving $\rho(P_{\delta^0}) < 1$. There is at least one. Suppose for some $\gamma \in \Delta$ that

$$H_\gamma \equiv \{v: A_{\delta^0} v \geqq c_{\delta^0}, A_\gamma v \leqq c_\gamma\}$$

is unbounded.

Thus there is a nonzero vector r such that $A_{\delta^0} r \geqq 0$ and $A_\gamma r \leqq 0$. Since $\rho(P_{\delta^0}) < 1$ and $P_{\delta^0} \geqq 0$, by Lemma 1, $A_{\delta^0}^{-1} \geqq 0$. Thus, $A_{\delta^0}^{-1} A_{\delta^0} r = r \geqq 0$.

Suppose $\rho(P_\gamma) < 1$. Then through a similar sequence of steps, we get $r \leqq 0$. Thus $r = 0$, a contradiction. Hence $\rho(P_\gamma) = 1$. By the Perron-Frobenius theorem there is a semipositive x, such that $x' A_\delta = 0$.

We now show that $x' c_\gamma = 0$. Since H_γ is unbounded there is a $v \in H_\gamma$. Then $0 = x' A_\gamma v \leqq x' c_\gamma$. Likewise, since D is nonempty by assumption, there is a $w \in D$ so that $A_\gamma w \geqq c_\gamma$. Hence, $0 = x' A_\gamma w \geqq x' c_\gamma$. Collecting results gives $x' c_\gamma = 0$. Note then that $x_i > 0$ implies $(A_\gamma w)_i = (c_\gamma)_i$ for each $w \in D$. This of course holds for $v^* \in D$.

In the following steps of the proof we will construct a policy $\beta \in \Delta$ such that $\rho(P_\beta) = 1$ and $A_\beta v = c_\beta$ has a solution which gives our desired contradiction.

As shown earlier, there is a semipositive r giving $A_{\delta^0} r \geqq 0$ and $A_\gamma r \leqq 0$. Partition and permute A_γ, x, and r as follows. First rearrange the terms so that all the zero values of x come first. Then rearrange terms so that the negative components of $A_\gamma r$ come first for the zero rows of x. Finally, rearrange terms so that the zero components

of r come first for the positive rows of x. This gives Table 1. Since x is semipositive, $N_3 \cup N_4$ is nonempty. Also, since $x'A_\gamma = 0$ and $x'A_\gamma r = 0$, rows N_3 and N_4 of $A_\gamma r \leqq 0$ must be zero-valued. We also see from $x'A_\gamma = 0$ that $P_{31} = 0$, $P_{32} = 0$, $P_{41} = 0$, and $P_{42} = 0$. From $A_\gamma r \leqq 0$ with equality in rows N_3, then $P_{34} = 0$. Rewriting the above using these observations gives Table 2.

If N_4 is nonempty, then $\rho(P_{44}) = 1$ from the last row of $A_\gamma r$ and the Perron-Frobenius theorem. If N_4 is empty, then $\rho(P_{33}) = 1$ from the third partition of $x'A_\gamma$ and the Perron-Frobenius theorem.

Let δ^* be as given in Theorem 2. Form a new policy β from γ and δ^* by replacing corresponding entries of δ^* by the entries of γ corresponding to the rows listed in N_4 (N_3 if N_4 is empty). From the Perron-Frobenius theorem, $\rho(P_\beta) = 1$ (since $\rho(P_{44}) = 1$ or $\rho(P_{33}) = 1$, respectively). Also, since $A_{\delta^*} v^* = c_{\delta^*}$ and $(A_\gamma v^*)_i = (c_{\gamma^*})_i$ for $i \in N_3 \cup N_4$, then $A_\beta v^* = c_\beta$. This gives the desired contradiction. Thus H_γ is bounded and $C_2 \subseteq C_5$. □

The following shows that Theorem 9 is as far as we can go with C_5.

Remark 1. C_3 is not contained in C_5^+. This can be seen with a simple counter-example. Let

$$\Delta = \{\delta, \gamma\}$$

where

$$A_\delta = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}, \qquad c_\delta = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and

$$A_\gamma = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}, \qquad c_\gamma = \begin{bmatrix} 1 \\ -5 \end{bmatrix}.$$

Here

$$F = D = \left\{ \begin{bmatrix} 3 \\ 1 \end{bmatrix} \right\}$$

TABLE 1

Rows	x	A_γ	r	$A_\gamma r$
N_1	0	$I - P_{11} \quad -P_{12} \quad -P_{13} \quad -P_{14}$	$\geqq 0$	< 0
N_2	0	$-P_{21} \quad I - P_{22} \quad -P_{23} \quad -P_{24}$	$\geqq 0$	$= 0$
N_3	> 0	$-P_{31} \quad -P_{32} \quad I - P_{33} \quad -P_{34}$	$= 0$	$= ?$
N_4	> 0	$-P_{41} \quad -P_{42} \quad -P_{43} \quad I - P_{44}$	> 0	$= ?$

TABLE 2

Rows	x	A_γ	r	$A_\gamma r$
N_1	0	$I - P_{11} \quad -P_{12} \quad -P_{13} \quad -P_{14}$	$\geqq 0$	< 0
N_2	0	$-P_{21} \quad I - P_{22} \quad -P_{23} \quad -P_{24}$	$\geqq 0$	$= 0$
N_3	> 0	$0 \quad 0 \quad I - P_{33} \quad 0$	$= 0$	$= 0$
N_4	> 0	$0 \quad 0 \quad -P_{43} \quad I - P_{44}$	> 0	$= 0$

and

$$\rho(P_\delta) = 0 < 1 \text{ yet } \{v: A_\delta v \geqq c_\delta, A_\gamma v \leqq c_\gamma\} = \left\{ \begin{bmatrix} 3 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \lambda : \lambda \geqq 0 \right\}$$

and

$$\{v: A_\gamma v \geqq c_\gamma, A_\delta v \leqq c_\delta\} = \left\{ \begin{bmatrix} 3 \\ 1 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} \lambda : \lambda \geqq 0 \right\}.$$

This observation suggests that there is a more general setting for Markov-type decisions than the conditions used by Eaves.

The following results cover classes C_4^+ and C_5^+.

Remark 2. C_2 is not a subset of C_4^+ since C_2 may have singular A_δ matrices but C_4^+ may not. This can be seen by the following counterexample. Let

$$\Delta = \{\delta, \gamma\}$$

where

$$A_\delta = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad c_\delta = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

and

$$A_\gamma = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}, \qquad c_\gamma = \begin{bmatrix} 2 \\ -4 \end{bmatrix}.$$

It is easy to verify that this problem belongs to C_2 but A_δ is singular.

C_1 is not equal to C_4^+. This also can be seen with a simple counterexample. Let

$$\Delta = \{\delta\}$$

where

$$A_\delta = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}, \qquad c_\delta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Here $F = \{0\}$ and $D = \{v: v \leqq 0\}$ but $\rho(P_\delta) > 1$.

This last example also shows that C_5^+ is not contained in C_1, C_2, or C_3.

The following result gives a relationship between C_1 and C_4^+.

THEOREM 10. *If* $(\Delta, A, c) \in C_4^+$ *and* $\rho(P_\delta) < 1$ *for some* $\delta \in \Delta$, *then* $(\Delta, A, c) \in C_1$.

Proof. Let $(\Delta, A, c) \in C_4^+$. Since every A_γ is nonsingular for $\gamma \in \Delta$, then $\rho(P_\gamma)$ is not equal to one. Suppose there exist $\gamma \in \Delta$ such that $\rho(P_\gamma) > 1$. Let $\delta_0, \delta_1, \cdots, \delta_m \in \Delta$ be formed from δ and γ with $\delta_0 = \delta$ and $\delta_m = \gamma$ and δ_j formed from δ_{j-1} by replacing the jth entry of δ_{j-1} by the jth entry of γ. Let k be the first integer such that $\rho(P_{\delta_k}) > 1$. Note that $k \geqq 1$.

We have $\rho(P_{\delta_{k-1}}) < 1$ and $\rho(P_{\delta_k}) > 1$. (Recall that no policy can have a spectral radius of one.) $P_{\delta_{k-1}}$ and P_{δ_k} differ only in their kth row. Let $\lambda \in (0, 1)$ and $P_\lambda = \lambda P_{\delta_{k-1}} + (1-\lambda) P_{\delta_k}$. By Lemma 3.7 from Seneta [15] (relaxed to reducible matrices) and the continuity properties of convex functions, we have that $\rho(P_\lambda) = 1$ for some $\lambda \in (0, 1)$. This contradicts our assumption that $(\Delta, A, c) \in C_4^+$ since all convex combinations of the A_δ's must be nonsingular. Hence $\rho(P_\gamma) < 1$ for each $\gamma \in \Delta$ and $(\Delta, A, c) \in C_1$. □

5. Additional problem classes. A number of other problem classes have been studied in the literature [6], [9], [13], [14], [16]. Hordijk and Kallenberg [6] have

shown the relationships between these. Below we incorporate their results and relate them to our results.

DEFINITION 10 (class *ST*—stochastic). Class *ST* is the class of all problems (Δ, A, c) satisfying the following conditions:

(1) $P_\delta \geqq 0$, for all $\delta \in \Delta$.

(2) $P_\delta \mathbf{1} = \mathbf{1}$, for all $\delta \in \Delta$.

(Here $\mathbf{1}$ is a vector of ones.)

DEFINITION 11 (class *SST*—substochastic). Class *SST* is the class of all problems (Δ, A, c) satisfying the following conditions:

(1) $P_\delta \geqq 0$, for all $\delta \in \Delta$.

(2) $P_\delta \mathbf{1} \leqq \mathbf{1}$, for all $\delta \in \Delta$.

DEFINITION 12 (class *DC*—discounted). Class *DC* is the class of all problems (Δ, A, c) satisfying the following conditions:

(1) $P_\delta \geqq 0$, for all $\delta \in \Delta$.

(2) $P_\delta \mathbf{1} = a\mathbf{1}$, for all $\delta \in \Delta$ and for some $a \in [0, 1)$ and fixed.

Clearly, $ST \subseteq SST$ and $DC \subseteq SST$. It has already been pointed out that $DC \subseteq C_1$.

DEFINITION 13 (class *E*—excessive). Class *E* is the class of all problems (Δ, A, c) satisfying the following conditions:

(1) $P_\delta \geqq 0$, for all $\delta \in \Delta$.

(2) $A\mathbf{u} \geqq 0$, for some $\mathbf{u} > 0$.

Clearly, $STT \subseteq E$ since $\mathbf{u} = \mathbf{1}$ provides the necessary vector. $C_1 \subseteq E$ is also easy to show [6].

Remark 3. $C_2 \subseteq E$. This follows from Corollary 5.2 in [11].

E is not contained in C_2 or C_4^+ as seen by the following simple counterexample. Let

$$\Delta = \{\delta\}$$

where

$$A_\delta = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = A, \qquad c_\delta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Clearly, $A\mathbf{1} = 0$ so $(\Delta, A, c) \in E$, but no $\delta \in \Delta$ gives $\rho(P_\delta) < 1$ and A_δ is singular.

Finally, C_4^+ is not contained in *E* as seen from the last example of Remark 2. There, no $\mathbf{u} > 0$ gives $A_\delta \mathbf{u} \geqq 0$.

Remark 4. C_3 is not contained in *E* as shown by the example used in Remark 1. The only $\mathbf{u}$ giving $A_\gamma \mathbf{u} \geqq 0$ is $\mathbf{u} = 0$.

DEFINITION 14 (class *N*—normalized). Class *N* is the class of all problems (Δ, A, c) satisfying the following conditions:

(1) $P_\delta \geqq 0$, for all $\delta \in \Delta$.

(2) $\rho(P_\delta) \leqq 1$, for all $\delta \in \Delta$.

Clearly, $E \subseteq N$ from the Perron-Frobenius theorem and the existence of *u*. Also, from Remark 3, $C_2 \subseteq N$. Neither C_3 nor C_4^+ is contained in *N* since both classes permit a spectral radius greater than one. Conversely, *N* is not contained in C_3 or C_4^+ for the same reasons discussed in Remarks 3 and 4 concerning *E*.

6. Summary. We have shown the following relationships:

(1) $C_1 \subseteq C_2 \subseteq C_3$.

(2) $C_4 \subseteq C_5$. $(C_4^+ \subseteq C_5^+.)$

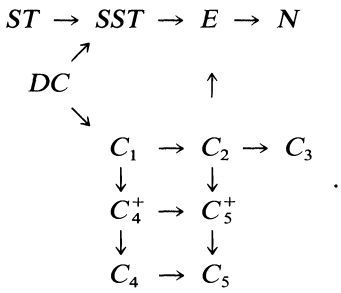
(3) $C_1 \subseteq C_4^+ \subseteq C_4$.

(4) $C_2 \subseteq C_5^+ \subseteq C_5$.

(5) $(\Delta, A, c) \in C_4^+$ and $\rho(P_\delta) < 1$ for some $\delta \in \Delta$ implies $(\Delta, A, c) \in C_1$.

(6) $C_2 \subseteq$ E.

The class memberships can be seen from the following diagram:



In addition, the following "negative" results hold:

(1) C_2 is not contained in C_4^+ and thus is not contained in C_4.
(2) C_3 is not contained in C_5^+ and thus is not contained in C_5.
(3) C_1 does not equal C_4^+.
(4) Neither C_4^+ nor C_5^+ is contained in C_1, C_2, or C_3.
(5) Neither E nor N is contained in C_2 or C_4^+.
(6) Neither C_3 nor C_4^+ is contained in E or N.

The results suggest that there might be a larger class of Markov-type decision problems than $C_4^+ \cup C_2$ where F contains a unique fixed point. The class of fixed-point problems of the form $v = L(v)$ with F nonempty is also probably larger than $C_3 \cup C_5$.

REFERENCES

[1] J. BATHER, *Optimal decision procedures for finite Markov chains. Part* III*: general convex systems*, Adv. in Appl. Probab., 5 (1973), pp. 541-553.
[2] R. W. COTTLE AND A. F. VEINOTT, *Polyhedral sets having a least element*, Math. Programming, 3 (1972), pp. 238-249.
[3] E. V. DENARDO, *Contraction mappings in the theory underlying dynamic programming*, SIAM Rev., 9 (1967), pp. 165-177.
[4] B. C. EAVES, *Complementary pivot theory and Markovian decision chains*, in Fixed Points: Algorithms and Applications, S. Karamardian, ed., Academic Press, New York, 1977, pp. 59-85.
[5] M. FIEDLER AND V. PTAK, *On matrices with non-positive off-diagonal elements and positive principal minors*, Czech. Math. J., 12 (1962), pp. 382-400.
[6] A. HORDIJK AND L. C. M. KALLENBERG, *Transient policies in discrete dynamic programming: linear programming including suboptimality tests and additional constraints*, Math. Programming, 30 (1984), pp. 46-70.
[7] R. A. HOWARD, *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, 1960.
[8] W. JEWELL, *Markov renewal programming* I *and* II, Oper. Res., 2 (1963), pp. 938-971.
[9] L. C. M. KALLENBERG, *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tract no. 148, Mathematical Centre, Amsterdam, 1980, Chap. 3.
[10] G. J. KOEHLER, *A generalized Markov decision process*, RAIRO Rech. Opér., 14 (1980), pp. 349-354.
[11] ———, *Value convergence in a generalized Markov decision process*, SIAM J. Control Optim., 17 (1979), pp. 180-186.
[12] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969, p. 34.
[13] U. G. ROTHBLUM, *Normalized Markov decision chains.* I*: sensitive discount optimality*, Oper. Res., 23 (1975), pp. 785-795.
[14] ———, *Optimality of non-stationary policies*, SIAM J. Control Optim., 15 (1977), pp. 221-232.
[15] E. SENETA, *Non-negative Matrices*, John Wiley, New York, 1973.
[16] K. M. VAN HEE, A. HORDIJK, AND J. VAN DER WAL, *Successive approximations for convergent dynamic programming*, in Markov Decision Theory, H. C. Tijms and J. Wessels, eds., Mathematical Centre Tract no. 93, Mathematical Centre, Amsterdam, 1977, pp. 183-211.
[17] A. F. VEINOTT, ED., *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist., 40 (1969), pp. 1635-1660.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1461-1480, November 1990

011

COMPUTABLE BOUNDS FOR THE SENSITIVITY OF THE ALGEBRAIC RICCATI EQUATION*

P. GAHINET† AND A. J. LAUB†

Abstract. In control or estimation theory, linear-quadratic optimization problems give rise to the so-called matrix algebraic Riccati equation (ARE). For such problems, a crucial issue is the existence and uniqueness of a symmetric nonnegative definite stabilizing solution to the ARE, and conditions on the equation parameters are known which guarantee both. However, in the context of computations in finite precision arithmetic, and with imperfect parameter identification, it is of concern whether the ARE retains such a solution in the proximity of a given set of parameters, and how sensitive this solution is to parameter variation.

In this paper, topological properties, such as openness of the domain of existence and continuity with respect to parameters, are established for the symmetric nonnegative definite stabilizing solution. Computable sensitivity estimates are also derived, which quantitatively define a region of safe computation, in terms of the parameters of the equation.

Key words. Riccati equation, sensitivity, stabilizability, computable bounds

AMS(MOS) subject classifications. 49E30, 93B35, 93B40

1. Introduction.

The symmetric algebraic Riccati equation (ARE) arises frequently in control and estimation problems. Consider the continuous-time ARE given by:

$$A^TX + XA - XFX + G = 0 \tag{1.1}$$

where all terms are matrices in $\mathbf{R}^{n\times n}$ (real square matrices of order n), and F and G are symmetric, nonnegative definite. The case of complex-valued matrices is qualitatively similar to the sequel but only the real-valued case will be considered here since it is most commonly encountered in applications. Under the assumption that the pairs (A, F) and (G, A) are stabilizable and detectable, respectively, there is a unique nonnegative definite symmetric stabilizing solution X to (1.1) (see [3] or [12]). By X stabilizing (for the pair (A, F)), we mean that $A - FX$ is stable, i.e., all its eigenvalues have strictly negative real parts.

Numerical algorithms are now available that solve the ARE efficiently and dependably, provided the original problem is sufficiently well-conditioned (see [13] or [1]). Well-conditioned means that the solution X is not greatly affected by small perturbations of the data A, F, G. In that case, and with an appropriate scaling of the data (cf. [8]), the Schur-type solvers yield accurate solutions to (1.1).

A natural question following this preliminary remark is how to assess the conditioning of the symmetric ARE, that is, its sensitivity to perturbations of the data. In other words, if we consider a perturbed version of (1.1):

$$(A+\Delta A)^TS + S(A+\Delta A) - S(F+\Delta F)S + G + \Delta G = 0, \tag{1.2}$$

under what conditions does (1.2) keep a unique, nonnegative definite stabilizing solution S? And can we estimate the maximum discrepancy $\|X - S\|$ for a given range of data perturbations ΔA, ΔF, ΔG?

* Received by the editors April 17, 1989; accepted for publication (in revised form) December 13, 1989.

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106. This research was supported by National Science Foundation grant ECS87-18897 and Air Force Office of Scientific Research contract AFOSR-89-0167.

In the first part of the paper, existing contributions to this problem will be reviewed. New results about the topological properties of the nonnegative definite stabilizing solution to (1.1) are then presented, and an existence condition for S, along with bounds for the variation $X-S$, are given in terms of the data perturbation. These bounds are shown to have a computable expression, that is, an expression that involves only known entities such as A, F, G, and some computed solution S_0. Finally, an application of this result to the Newton refinement of solutions to (1.1) (see [9]) is discussed.

The following notation and definitions are used in what follows.

(1) The space of linear operators over $\mathbf{R}^{n\times n}$ will be denoted by $\mathbf{L}$.

(2) Throughout the paper, the vector norm will be the Euclidean norm, and the matrix norm the spectral norm defined as

$$\text{for } M\in\mathbf{R}^{n\times n},\quad \|M\|=\sup_{\substack{z\in\mathbf{R}^n\\ z\neq 0}}\frac{\|Mz\|}{\|z\|}.$$

(3) The spectral norm will also be used over the space $\mathbf{L}$; that is,

$$\text{for } \Omega\in\mathbf{L},\quad \|\Omega\|=\sup_{\substack{X\in\mathbf{R}^{n\times n}\\ X\neq 0}}\frac{\|\Omega(X)\|}{\|X\|}.$$

(4) The subset of $\mathbf{R}^{n\times n}$ consisting of the real symmetric matrices will be denoted by $\mathbf{H}_n$, and the subset of $\mathbf{H}_n$ consisting of the symmetric nonnegative definite matrices by $\mathbf{H}_n^+$.

2. A survey on sensitivity estimates. A natural way to analyze the sensitivity of (1.1) is to look at the maximum perturbation of X (in relative terms) resulting from data perturbations of a given magnitude. The perturbations of the data A, F, G will be restricted to those perturbations preserving the nonnegative definiteness of the matrices F and G. This constraint may appear impractical in the context of random perturbations, but recall that the matrices F and G arise naturally in factored form. Specifically, for a system whose dynamics and output are described by

$$\dot{x}=Ax+Bu;\qquad A\in\mathbf{R}^{n\times n};\qquad B\in\mathbf{R}^{n\times m};$$

$$y=Cx;\qquad C\in\mathbf{R}^{p\times n},$$

we will have $F=BB^T$ and $G=C^TC$ in the associated ARE. Thus, even when errors in B or C are random, F and G still remain in $\mathbf{H}_n^+$.

For $\delta>0$, consider the set of perturbations P_δ defined by

(2.1)
$$P_\delta=\left\{(\Delta A,\Delta F,\Delta G): F+\Delta F,\ G+\Delta G \text{ in } \mathbf{H}_n^+ \text{ and } \max\left(\frac{\|\Delta A\|}{\|A\|},\frac{\|\Delta F\|}{\|F\|},\frac{\|\Delta G\|}{\|G\|}\right)\leqq\delta\right\}.$$

Note that P_δ is convex, and hence connected. The number δ will be called the magnitude of perturbation. The sensitivity of (1.1) to perturbations in P_δ is then measured by

$$K_\delta=\sup\left\{\frac{\|\Delta X\|}{\delta\|X\|}: (\Delta A,\Delta F,\Delta G)\in P_\delta\right\}. \tag{2.2}$$

In most applications we are only interested in small data perturbations. This motivates the definition by Rice [15] of a condition number K for the ARE (1.1), obtained as the limit when $\delta\to 0$ of the K_δ's.

The condition number K can be closely approximated by another sensitivity estimate, due to Byers [2]. Consider the following linear operators:

$$\begin{aligned}\Omega_X(Z) &= (A-FX)^T Z + Z(A-FX)\\ \Theta_X(Z) &= \Omega_X^{-1}(Z^T X + XZ)\\ \Pi_X(Z) &= \Omega_X^{-1}(XZX),\end{aligned} \tag{2.3}$$

and the number $K_B(X)$ defined as

$$K_B(X) = \frac{\|\Omega_X^{-1}\|\,\|G\| + \|\Theta_X\|\,\|A\| + \|\Pi_X\|\,\|F\|}{\|X\|}. \tag{2.4}$$

Then K and K_B are related by $\frac{1}{3}K_B \leqq K \leqq K_B$ [10]. The estimate $K_B(X)$ is essentially an a posteriori estimate since it involves the solution X. This may seem a serious drawback, since we can only hope to obtain an approximate solution S when solving (1.1) numerically. Nevertheless, replacing X by S in the expression of K_B is technically possible, as we will see in § 5. Also, note that the norms $\|\Omega_S^{-1}\|$, $\|\Theta_S\|$, $\|\Pi_S\|$ can be easily computed or approximated (see Theorem 2.4 in [10]). For instance, $\|\Omega_S^{-1}\|$ is equal to the norm of the solution H to

$$(A-FS)^T H + H(A-FS) + I = 0.$$

The approximate condition number K_B does not offer a totally satisfying answer to the sensitivity issue, however, since it corresponds to a limiting case where the magnitude of perturbation goes to zero. It merely indicates that, to first order in δ,

$$\frac{\|\Delta X\|}{\|X\|} \approx K\delta,$$

with K bounded by K_B. Making explicit the contribution of the higher order terms in δ is one key motivation of this paper.

Finally, another class of results should be mentioned here: the so-called residual bounds. Although they do not address the general sensitivity problem as formulated above, they are of great practical interest in estimating the accuracy of an approximate solution S to (1.1) (typically a computed solution). Using the residual to bound the error in the solution is a common technique in linear operator theory. If Ψ is an invertible linear operator, X solves $\Psi(X) = Q$, and S approximates X, then the residual is by definition $R = \Psi(S) - Q$, and the error in the solution is bounded by

$$\|X - S\| \leqq \|\Psi^{-1}\|\,\|R\|.$$

This has a straightforward application to finite systems of linear equations $Ax = b$ ($\Psi = A$), and, although the operator associated with the ARE is not linear, an analogous manipulation on its linear part leads to the following result [9].

THEOREM 2.1. *Let X be the exact solution of* (1.1), *and suppose there is a stabilizing solution S to* (1.2). *Let $R = G + A^T S + SA - SFS$ be the residual. If $4\|F\|\,\|\Omega_S^{-1}\|^2\,\|R\| < 1$ and $\|X - S\| < 1/(3\|F\|\,\|\Omega_S^{-1}\|)$, then*

$$\|X - S\| \leqq 2\|\Omega_S^{-1}\|\,\|R\|. \tag{2.5}$$

A problem with this theorem is the checkability of the second condition, which involves the unknown X. This drawback was avoided in a more specialized result, which considers the particular problem of estimating the error between X and the approximate solution computed by a Schur-type algorithm (see [8]). If the matrix S in Theorem 2.1 comes from such computations, then the second condition can be

checked through the estimate in [8]. Note that the estimation of algorithm-dependent errors fits in the general frame of sensitivity analysis; that is, the computed solution can be viewed as the solution of a perturbed problem (1.2), provided the algorithm is stable. Nevertheless, the magnitude of data perturbation involved is very difficult to assess. On the other hand, the bound proposed in this paper is algorithm-independent, but requires that the magnitude of data perturbation be known. The two results are therefore complementary: combined, both make the second condition of Theorem 2.1 checkable in most situations.

3. Topological properties of the ARE solution. In many applications, it is desirable to have a unique symmetric nonnegative definite stabilizing (USNDS) solution to (1.1). Criteria are available to establish the existence of such a solution for a particular nominal set of parameters A, F, G, but, once established at a nominal set, can we conclude anything for nearby parameter sets? In this section, theoretical results are established which show the existence and continuity of the USNDS in a small neighborhood of a parameter set A, F, G for which there is a USNDS solution to (1.1). The appropriate framework here is the complete metric space $(\mathbf{T}, d)$, where $\mathbf{T}$ and the metric d are defined by

$$\mathbf{T} = \mathbf{R}^{n\times n} \times \boldsymbol{H}_n \times \mathbf{H}_n = \{(A, F, G): A \text{ in } \mathbf{R}^{n\times n} \text{ and } F, G \text{ in } \mathbf{H}_n\},$$

$$d((A_1, F_1, G_1), (A_2, F_2, G_2)) = \|A_1 - A_2\| + \|F_1 - F_2\| + \|G_1 - G_2\|.$$

Finally, to simplify description, we will call "USNDS parameter set" any parameter triple (A, F, G) in $\mathbf{T}$ for which (1.1) has a USNDS solution.

LEMMA 3.1. *Suppose* (1.1) *has a* USNDS *solution* X. *If the perturbation* $(\Delta A, \Delta F, \Delta G)$ *is such that* $F+\Delta F$ *and* $G+\Delta G$ *are in* $\mathbf{H}_n^+$, *and also satisfies*

$$\|\Delta A\| + \|\Delta F\|\,\|X\| < \frac{1}{2\|\Omega_X^{-1}\|}, \tag{3.1}$$

then (1.2) *has a unique symmetric nonnegative definite solution* S, *such that* $A - FS$ *has all its eigenvalues in the closed complex left half-plane. Such a solution is referred to as a* strong *solution . Furthermore, if*

$$\|X - S\| < \frac{1}{2\|F\|\,\|\Omega_X^{-1}\|}, \tag{3.2}$$

then S *is stabilizing for the pair* (A, F), *and if*

$$\|\Delta A\| + \|\Delta F\|\,\|X\| < \frac{1}{4\|\Omega_X^{-1}\|} \quad \text{and} \quad \|X - S\| < \frac{1}{4\|F+\Delta F\|\,\|\Omega_X^{-1}\|}, \tag{3.3}$$

then S *is the* USNDS *solution to* (1.2).

Proof. This theorem relies on the following result (Theorem 1.2 in [7]). Given a matrix R, and the associated Lyapunov operator $\Gamma(Z) := R^T Z + ZR$, if R is stable and $\|\Delta R\| < 1/(2\|\Gamma^{-1}\|)$, then $R+\Delta R$ is also stable.

From the definition of X, the matrix $A - FX$ is stable. Now,

$$(A+\Delta A) - (F+\Delta F)X = A - FX + \Delta A - \Delta FX = A - FX + E,$$

with $\|E\| \leqq \|\Delta A\| + \|\Delta F\|\,\|X\|$. The assumption (3.1) guarantees that

$$\|E\| < \frac{1}{2\|\Omega_X^{-1}\|},$$

and the stability of the matrix $(A+\Delta A)-(F+\Delta F)X$ follows by applying the result in [7] to $A-FX$. Consequently, the pair $(A+\Delta A, F+\Delta F)$ is stabilizable and Theorem 1 in [14] in turn guarantees the existence of a unique strong solution in the sense of Chan, Goodwin, and Sin [4].

To show that S is stabilizing for (A, F) when (3.2) holds, write $A-FS = A-FX+F(X-S)$. Invoking again the result in [7], we see that $A-FS$ will be stable whenever $\|F\|\,\|X-S\| < 1/(2\|\Omega_X^{-1}\|)$, which is the condition (3.2). Finally, by writing

$$(A+\Delta A)-(F+\Delta F)S = A-FX+\Delta A-\Delta FX-(F+\Delta F)\Delta X,$$

it follows from (3.3) that the norm of the perturbation of $A-FX$ in the right-hand side above is again less than $1/(2\|\Omega_X^{-1}\|)$, whence the stability of $(A+\Delta A)-(F+\Delta F)S$. □

A fundamental result regarding the regularity of the USNDS solution to (1.1) can be found in [5] and is recalled in the next theorem. For the sake of completeness, the proof is also included.

THEOREM 3.2. *The set of parameter triples* (A, F, G) *for which* (1.1) *has a* USNDS *solution is an open set in the metric space* $(\mathbf{T}, d)$. *That is, if* (1.1) *has a* USNDS *solution* X_0 *for some parameters* (A_0, F_0, G_0) *(in* $\mathbf{T}$*), then there is some* $\varepsilon > 0$ *such that* (1.1) *has a* USNDS *solution for any parameter set* (A, F, G) *in* $\mathbf{T}$ *within* ε *of* (A_0, F_0, G_0) *for the metric d. Moreover, in this open set, the* USNDS *solution depends continuously on, and is, in fact, infinitely differentiable with respect to the parameters.*

Proof. Let (A_0, F_0, G_0) be a set of parameters for which (1.1) has a USNDS solution X_0. Consider the matrix-valued functional

$$f: \mathbf{H}_n \times (\mathbf{R}^{n\times n} \times \mathbf{H}_n \times \mathbf{H}_n) \to \mathbf{H}_n$$

defined by

$$f(X, A, F, G) = A^T X + XA - XFX + G.$$

By assumption, $f(X_0, A_0, F_0, G_0) = 0$. As a quadratic function in X, and a linear function in A, F, and G, the function f is differentiable (even infinitely differentiable), and its derivative with respect to X at the point (X_0, A_0, F_0, G_0) is the linear operator given for any matrix Z by

$$Df_X(Z) = (A_0 - F_0X_0)^T Z + Z(A_0 - F_0X_0).$$

Since X_0 is stabilizing, the operator Df_X is nonsingular. Therefore, from the Implicit Function Theorem (see, e.g., [16, p. 356]), there exist

- open neighborhoods U and V of (X_0, A_0, F_0, G_0) and (A_0, F_0, G_0), respectively, with $U \subset \mathbf{H}_n \times \mathbf{T}$ and $V \subset \mathbf{T}$,
- an infinitely differentiable matrix-valued function Ψ defined in V,

such that, for any (A, F, G) in V, $X = \Psi(A, F, G)$ is the only solution to

$$f(X, A, F, G) = 0 \quad \text{and} \quad (X, A, F, G) \in U. \tag{3.4}$$

Now, by continuity of Ψ, we can take V small enough so that for (A, F, G) in V and $X = \Psi(A, F, G)$,

$$\|(A-FX)-(A_0-F_0X_0)\| < \frac{1}{2\|\Omega_{X_0}^{-1}\|},$$

which then guarantees that $A-FX$ is stable. Note that (1.1) can be rewritten

$$(A-FX)^T X + X(A-FX) + G + XFX = 0.$$

Since $G+XFX$ is nonnegative definite and $A-FX$ is stable, X itself must be nonnegative definite from the Lyapunov Theorem. Therefore, for any (A, F, G) in a small enough open neighborhood V of (A_0, F_0, G_0), (1.1) has a USNDS solution $X=\Psi(A, F, G)$. The uniqueness is a consequence of the stabilizability of (A, F) (see [14]). This completes the proof of the openness in $(\mathbf{T}, d)$ of the set of USNDS parameter triples.

The continuity and infinite differentiability of the USNDS solution as a function of the parameters (A, F, G) follows immediately from the properties of Ψ as stated above. □

We call a curve of USDNS solutions any curve of points (X, A, F, G) such that X is the USNDS solution to (1.1) for the parameters (A, F, G).

COROLLARY 3.3. *The curves of* USNDS *solutions are isolated, in the sense that in a neighborhood of such curves, there is no other quadruple* (X', A', F', G') *solving* $f(X', A', F', G')=0$ (*f as in Theorem* 3.2).

Proof. This result is contained in Theorem 3.2. Specifically, consider a curve of USNDS solutions passing through the point (X_0, A_0, F_0, G_0). Then there is an open neighborhood of this point within which all the solutions (X, A, F, G) to (3.4) are exclusively USNDS; that is, the curve of USNDS solutions is isolated. □

The continuous dependence on the parameters is not limited to the USNDS solution; given a USNDS parameter set (A_0, F_0, G_0), it indeed extends to the strong solution defined for all parameter sets in

$$\mathbf{D}=\left\{(A_0+\Delta A, F_0+\Delta F, G_0+\Delta G)\in\mathbf{T}:\right.$$

$$\left.F_0+\Delta F, G_0+\Delta G \text{ in } \mathbf{H}_n^+ \text{ and } \|\Delta A\|+\|\Delta F\|\,\|X_0\|<\frac{1}{2\|\Omega_{X_0}^{-1}\|}\right\}. \tag{3.5}$$

Note the constraint $F_0+\Delta F$ and $G_0+\Delta G$ nonnegative definite. To prove this result, the following lemma is needed first.

LEMMA 3.4. *Suppose* (1.1) *has a* USNDS *solution* X_0 *for some parameters* A_0, F_0, G_0, *and consider for* $\varepsilon>0$ *the compact subset of* D:

$$\boldsymbol{D}_\varepsilon=\left\{(A_0+\Delta A, F_0+\Delta F, G_0+\Delta G)\in\mathbf{D}:\right.$$

$$\left.\|\Delta A\|+\|\Delta F\|\,\|X_0\|\leqq\frac{1}{2\|\Omega_{X_0}^{-1}\|}-\varepsilon \text{ and } \|\Delta G\|\leqq\frac{1}{\varepsilon}\right\}.$$

Then, for any set of parameters (A, F, G) *in* $\mathbf{D}_\varepsilon$, (1.1) *has a unique strong solution. Moreover, this strong solution, as a function of* (A, F, G), *is uniformly bounded in* $\mathbf{D}_\varepsilon$.

Proof. The condition (3.1) of Lemma 3.1 is fulfilled for any (A, F, G) in $\mathbf{D}_\varepsilon$ since

$$\|A-A_0\|+\|F-F_0\|\,\|X_0\|\leqq\frac{1}{2\|\Omega_{X_0}^{-1}\|}-\varepsilon<\frac{1}{2\|\Omega_{X_0}^{-1}\|}.$$

The existence and uniqueness of a strong solution in $\mathbf{D}_\varepsilon$ is thus clear. Note that this strong solution coincides with the USNDS solution whenever the latter exists.

Suppose the strong solution is unbounded in $\mathbf{D}_\varepsilon$. Then there exists a sequence of parameters $\{A_k, F_k, G_k\}_{k=1}^\infty$ in $\mathbf{D}_\varepsilon$, such that the norm of the strong solution X_k to

$$A_k^TX_k+X_kA_k-X_kF_kX_k+G_k=0 \tag{3.6}$$

goes to infinity with k. For each k, let $0\leqq\lambda_n^{(k)}\leqq\cdots\leqq\lambda_1^{(k)}$ be the (real) eigenvalues of X_k, and $u_n^{(k)}, \cdots, u_1^{(k)}$ an associated orthonormal system of eigenvectors. Observe that the sequences $\{A_k, F_k, G_k\}_{k=1}^\infty$ and $\{u_i^{(k)}\}_{k=1}^\infty$ $(i=1, \cdots, n)$ all lie in compact sets and

thus have convergent subsequences. Up to a reindexing of these sequences, assume that they all actually converge, and let (A, F, G) and u_i $(i=1, \cdots, n)$ denote their respective limits. Note that (A, F, G) lies in $\mathbf{D}_\varepsilon$ (as the limit of points of the closed set $\mathbf{D}_\varepsilon$), and that the u_i's form an orthonormal basis.

Since $\|X_k\|$ goes to infinity with k, there is some integer r less than n, such that the sequence $\{\lambda_i^{(k)}\}_{k=1}^{\infty}$ is bounded for $i<r$, and unbounded otherwise. Premultiply (3.6) by $u_i^{(k)T}$ and postmultiply by $u_j^{(k)}$. This yields

$$\lambda_j^{(k)} u_i^{(k)T} A_k^T u_j^{(k)} + \lambda_i^{(k)} u_i^{(k)T} A_k u_j^{(k)} - \lambda_i^{(k)} \lambda_j^{(k)} u_i^{(k)T} F_k u_j^{(k)} + u_i^{(k)T} G_k u_j^{(k)} = 0. \tag{3.7}$$

First consider the case where both i and j are $\geqq r$. Then both $\lambda_i^{(k)}$ and $\lambda_j^{(k)}$ go to infinity as $k \to \infty$. Dividing (3.7) by $\lambda_i^{(k)} \lambda_j^{(k)}$ and letting k go to infinity, we then obtain

$$\lim_{k\to\infty} u_i^{(k)T} F_k u_j^{(k)} = u_i^T F u_j = 0 \quad \text{for all } i, j \geqq r. \tag{3.8}$$

Now consider the case $i \geqq r$ and $j < r$. Divide (3.7) by $\lambda_i^{(k)}$ and let $k \to \infty$. This yields

$$\lim_{k\to\infty} u_i^{(k)T} A_k u_j^{(k)} = u_i^T A u_j = 0 \quad \text{for all } i \geqq r \quad \text{and} \quad j < r. \tag{3.9}$$

Based on (3.8) and (3.9), the matrices A and F can be partitioned with respect to the basis $\{u_1, \cdots, u_n\}$ as:

$$A \equiv \begin{pmatrix} \hat{A} & * \\ 0 & \tilde{A} \end{pmatrix}; \qquad F \equiv \begin{pmatrix} \hat{F} & \Delta \\ \Delta^T & 0 \end{pmatrix}, \tag{3.10}$$

where the blocks $\hat{A}$ and $\hat{F}$ are $(r-1)\times(r-1)$. Moreover, the nonnegative definiteness of F imposes $\Delta = 0$. Now, since (A, F, G) is in $\mathbf{D}_\varepsilon$, (A, F) is stabilizable, which, combined with (3.10) and $\Delta = 0$, requires that $\tilde{A}$ be stable.

A contradiction to the initial assumption that "the strong solution is unbounded in $\mathbf{D}_\varepsilon$" can now be derived as follows. Consider the case $i = j \geqq r$ in (3.7), divide by $\lambda_i^{(k)}$ and let k tend to infinity. This yields

$$\lim_{k\to\infty} (2 u_i^{(k)T} A_k u_i^{(k)} - \lambda_i^{(k)} u_i^{(k)T} F_k u_i^{(k)}) = 0.$$

But $\lim_{k\to\infty} u_i^{(k)T} A_k u_i^{(k)} = u_i^T A u_i$ and therefore

$$u_i^T A u_i = \lim_{k\to\infty} \frac{\lambda_i^{(k)} u_i^{(k)T} F_k u_i^{(k)}}{2},$$

which implies that $u_i^T A u_i \geqq 0$ since $\lambda_i^{(k)} u_i^{(k)T} F_k u_i^{(k)} \geqq 0$ for all k. Consequently,

$$\text{Trace}\,(\tilde{A}) = \sum_{i=r}^{n} u_i^T A u_i \geqq 0,$$

which indeed contradicts the stability of $\tilde{A}$. □

The previous lemma is crucial to the following topological result about the strong solution of the ARE (1.1) for parameters in $\mathbf{D}$.

THEOREM 3.5. *Suppose* (1.1) *has a* USNDS *solution* X_0 *for some parameters* A_0, F_0, G_0. *Then the strong solution to nearby problems* (1.2), *considered as a function of the parameters* $A_0+\Delta A$, $F_0+\Delta F$, $G_0+\Delta G$, *is continuous over* $\mathbf{D}$ *as defined in* (3.5).

Proof. To prove the continuity of the strong solution in $\mathbf{D}$, consider a point (A, F, G) in $\mathbf{D}$, and a sequence of points $\{A_k, F_k, G_k\}_{k=1}^{\infty}$ in $\mathbf{D}$ converging to (A, F, G). Let X and X_k denote the corresponding (unique) strong solutions to (1.1) and (3.6), respectively.

By an appropriate choice of ε, we can ensure that all (A_k, F_k, G_k)'s and (A, F, G) lie in $\mathbf{D}_\varepsilon$ (use the openness of $\mathbf{D}$). Lemma 3.4 therefore applies to say that the sequence $\{X_k\}_{k=1}^{\infty}$ is bounded, and consequently has some convergent subsequence. Let X_∞ be the limit of such a subsequence.

Clearly, X_∞ solves the same ARE as X (and is also a strong solution) as the limit of a sequence of strong solutions. By the uniqueness of the strong solution in $\mathbf{D}$, X_∞ and X must be equal. Thus, the sequence $\{X_k\}$ can have only one accumulation point, X, and must converge to X. This proves the continuity of the strong solution at any point (A, F, G) in $\mathbf{D}$. □

4. Uniform bounds for the sensitivity of the USNDS solution. In the previous section, the continuity of the USNDS solution to (1.1) was qualitatively established. We now turn to quantitative estimates for the sensitivity of the USNDS solution to perturbations of the parameters. Throughout the section, (1.1) is assumed to have a USNDS solution X. The first theorem bounds the variation $\|X-S\|$ between X and a solution S to a nearby problem (1.2), in terms of A, F, G, X and the magnitude of perturbation δ. In the two results following, X is then replaced by a known approximation S_0 in order to make this bound computable.

The following lemma is needed in the proof of the first theorem.

LEMMA 4.1. *Let M_X and N_X be defined by*

$$M_X = (\|A\| + \|F\|\,\|X\|)\|\Omega_X^{-1}\|, \qquad N_X = K_B(X)\|X\|\,\|F\|\,\|\Omega_X^{-1}\|, \tag{4.1}$$

and a, b, c by

$$a = \|\Omega_X^{-1}\|\,\|F\|(1+\delta), \qquad 2b = 1-2M_X\delta, \qquad c = K_B(X)\|X\|\delta. \tag{4.2}$$

If

$$0 \leqq \delta < \frac{1}{1+4(M_X+N_X)}, \tag{4.3}$$

then

$$b^2 - ac > 0.$$

Proof. Assume (4.3). For the sake of clarity, the subscript X will be dropped from M_X and N_X in this proof. Noting that $ac = N\delta(1+\delta)$, $b^2 > ac$ is equivalent to

$$4(M^2-N)\delta^2 - 4(M+N)\delta + 1 > 0. \tag{4.4}$$

If $M^2 \geqq N$, a sufficient condition for (4.4) to hold is $1-4(M+N)\delta > 0$. This condition is fulfilled with the assumption (4.3), and therefore $b^2 > ac$ in this case.

In the other case ($M^2 < N$), consider the quadratic function

$$h(\delta) = 4(M^2-N)\delta^2 - 4(M+N)\delta + 1.$$

Its value is 1 at $\delta = 0$, and it goes to $-\infty$ when $|\delta| \to +\infty$, since $M^2 - N < 0$. Therefore, h will be strictly positive on any interval $[0, \hat{\delta}]$ such that $\hat{\delta} > 0$ and $h(\hat{\delta}) > 0$. But now,

$$h\left(\frac{1}{1+4(M+N)}\right) = \frac{(2M+1)^2}{(1+4(M+N))^2} > 0.$$

Hence, $h(\delta)$ is strictly positive for δ in $[0, 1/(1+4(M+N))]$, and (4.4) holds in the case $M^2 < N$ as well. This proves the lemma for any strictly positive M, N. □

THEOREM 4.2. *Assume* (1.1) *has a* USNDS *solution* X, *and let* M_X *and* N_X *be defined as in* (4.1). *Then, for perturbations* ΔA, ΔF, ΔG *in* P_δ *with*

$$\delta < \delta_0 = \frac{1}{4+8(M_X+N_X)}, \tag{4.5}$$

the equation (1.2) *has a* USNDS *solution* S, *also stabilizing for* (A, F), *and such that*

$$\|X-S\| \leqq \frac{K_B(X)\|X\|\delta}{1-3(M_X+N_X)\delta}, \tag{4.6}$$

where $K_B(X)$ *is given by* (2.4).

Proof. Again for clarity, K_B, M, and N will be used in place of $K_B(X)$, M_X, and N_X, respectively, in the proof. Assume δ satisfies (4.5), and consider ΔA, ΔF, ΔG in P_δ. Observe that

$$\begin{aligned}\|\Delta A\|+\|\Delta F\|\,\|X\| &\leqq (\|A\|+\|F\|\,\|X\|)\delta \\ &\leqq (\|A\|+\|F\|\,\|X\|)\frac{1}{4M} = \frac{1}{4\|\Omega_X^{-1}\|}.\end{aligned}$$

Thus, by Lemma 3.1, there is a unique strong solution $S = X + \Delta X$ to (1.2). To establish (4.6), first subtract (1.1) from (1.2) to obtain

$$\begin{aligned}\Delta X = &-\Omega_X^{-1}(\Delta G+\Delta A^T X + X\Delta A - X\Delta F X) \\ &-\Omega_X^{-1}((\Delta A-\Delta F X)^T\Delta X + \Delta X(\Delta A-\Delta F X)) \\ &+\Omega_X^{-1}(\Delta X(F+\Delta F)\Delta X).\end{aligned}$$

From the definition of K_B and δ, we can bound the norm of the first term on the right-hand side by $K_B\|X\|\delta$. The second and third terms are easily bounded to produce

$$\begin{aligned}\|\Delta X\| \leqq K_B\|X\|\delta + 2\|\Omega_X^{-1}\|(\|A\|+\|F\|\,\|X\|)\delta\|\Delta X\| \\ +\|\Omega_X^{-1}\|\,\|F\|(1+\delta)\|\Delta X\|^2.\end{aligned}$$

With a, b, c defined as in (4.2), this inequality becomes

$$a\|\Delta X\|^2 - 2b\|\Delta X\| + c = a\left(\|\Delta X\| - \frac{b}{a}\right)^2 + c - \frac{b^2}{a} \geqq 0. \tag{4.7}$$

With assumption (4.5), the condition of Lemma 4.1 is satisfied and therefore $b^2 - ac > 0$. It follows from (4.7) that either

$$\|\Delta X\| \leqq r_1(\delta) = \frac{b-\sqrt{b^2-ac}}{a}, \tag{4.8}$$

or

$$\|\Delta X\| \geqq r_2(\delta) = \frac{b+\sqrt{b^2-ac}}{a}. \tag{4.9}$$

The inequality (4.8) will provide the desired bound. Now we claim that (4.9) cannot hold for any perturbation in P_δ and δ satisfying (4.5). First, observe that the strong solution S is a continuous function over P_δ, since it is continuous over $\mathbf{D}$ (Theorem 3.5) and $P_\delta \subset \mathbf{D}$ from (4.5). Therefore, the real-valued function

$$\nu(\Delta A, \Delta F, \Delta G) = \|S(A+\Delta A, F+\Delta F, G+\Delta G) - S(A, F, G)\| = \|S-X\|$$

is continuous over P_δ, and since P_δ is connected, the set $\nu(P_\delta)$ must also be connected in $\mathbf{R}$, that is, it must be an interval. If (4.9) held for some perturbation in P_δ, then this interval would be contained in $[r_2(\delta), +\infty)$, because $r_1(\delta) < r_2(\delta)$, and no value between these two numbers is in $\nu(P_\delta)$. This is obviously impossible, since $r_2(\delta) > 0$ and $0 \in \nu(P_\delta)$ $(0 = \nu(0, 0, 0))$.

Therefore, under assumption (4.5), (4.8) holds. To obtain the bound (4.6), it is sufficient to show that

$$r_1(\delta) \leqq \frac{K_B \|X\| \delta}{1 - 3(M+N)\delta}. \tag{4.10}$$

Since $ax^2 - 2bx + c < 0$ if and only if $r_1(\delta) < x < r_2(\delta)$, inequality (4.10) will in turn be satisfied if

$$a\left(\frac{K_B\|X\|\delta}{1-3(M+N)\delta}\right)^2 - 2b\left(\frac{K_B\|X\|\delta}{1-3(M+N)\delta}\right) + c < 0.$$

Now, $1 - 3(M+N)\delta \neq 0$ when (4.5) holds; hence the sign of the left-hand side expression above will be the same as that of

$$\begin{aligned} &a(K_B\|X\|\delta)^2 - 2bK_B\|X\|\delta(1-3(M+N)\delta) + c(1-3(M+N)\delta)^2 \\ &\quad = K_B\|X\|\delta[(1+\delta)N\delta - (1-2M\delta)(1-3(M+N)\delta) \\ &\qquad + (1-3(M+N)\delta)^2] \\ &\quad = K_B\|X\|\delta^2\left[(\delta-2)N - M + 6(M+N)\delta\left(\frac{M}{2}+\frac{3N}{2}\right)\right]. \end{aligned}$$

By virtue of (4.5), $6(M+N)\delta < 1$ and $\delta < 1/2$, which guarantees that

$$(\delta-2)N - M + 6(M+N)\delta\left(\frac{M}{2}+\frac{3N}{2}\right) < -\frac{3N}{2} - M + \frac{M}{2} + \frac{3N}{2} = -\frac{M}{2} < 0.$$

Thus, (4.10) holds under assumption (4.5), and (4.6) follows.

Finally, the criteria of Lemma 3.1 are used to establish the stability results. The solution X is stabilizing for (A, F). Observe that a common lower bound for the right-hand sides of both (3.2) and (3.3) is $1/(4(1+\delta)\|F\|\,\|\Omega_X^{-1}\|)$. Using the bound (4.6) a sufficient condition for (3.2) and (3.3) to hold appears to be

$$\frac{K_B\|X\|\delta}{1-3(M+N)\delta} \leqq \frac{1}{4(1+\delta)\|F\|\,\|\Omega_X^{-1}\|},$$

or equivalently,

$$4N(1+\delta)\delta \leqq 1 - 3(M+N)\delta.$$

Since $\delta < 1/4$, this will in turn be true when

$$5N\delta \leqq 1 - 3(M+N)\delta, \quad \text{that is,} \quad \delta \leqq \frac{1}{3M+8N}.$$

This last condition is clearly implied by (4.5), whence S is stabilizing for (A, F) and $(A+\Delta A, F+\Delta F)$. □

The bound in the previous theorem has the inconvenience of being formulated in terms of the solution X to (1.1), which in practice is never known exactly. This drawback is removed in the next two results, which offer bounds involving only *known* quantities, and thus actually *computable*, that is, checkable quantities from the available data.

Two different situations are considered. In the first scenario, only approximations $A+\Delta A_0$, $F+\Delta F_0$, $G+\Delta G_0$ of the data A, F, G are known, along with the (exact) solution S_0 of the corresponding problem (1.2). This is the case in particular when the data is measured with a nonnegligible level of inaccuracy. In the second scenario, the data is known exactly, but only an approximation S_0 of X is available, which solves a nearby problem (1.2) whose parameters are unknown. The limitations of computations in finite arithmetic, for instance, can be modeled this way.

THEOREM 4.3. *Assume that only approximate values* $A+\Delta A_0$, $F+\Delta F_0$, $G+\Delta G_0$ *of the parameters* A, F, G *of* (1.1) *are known. Assume also that the corresponding equation* (1.2) *has a* USNDS *solution* S_0, *which is known exactly. Finally, let* $\hat{K}_B(S_0)$, $\hat{M}_{S_0}$, *and* $\hat{N}_{S_0}$ *be the counterparts of* $K_B(X)$, M_X, *and* N_X, *when* $A+\Delta A_0$, $F+\Delta F_0$, $G+\Delta G_0$, *and* S_0 *replace* A, F, G, *and* S *respectively, in* (2.3), (2.4), *and* (4.1).

If the perturbation $(\Delta A_0, \Delta F_0, \Delta G_0)$ *is in* P_δ *with*

$$\delta < \delta_1 = \frac{1}{4+8(\hat{M}_{S_0}+\hat{N}_{S_0})}, \tag{4.11}$$

then the equation (1.1) *has a* USNDS *solution* X, *such that*

$$\|X-S_0\| \leqq \frac{\hat{K}_B(S_0)\|S_0\|\delta}{1-3(\hat{M}_{S_0}+\hat{N}_{S_0})\delta}. \tag{4.12}$$

Proof. This result is a direct consequence of Theorem 4.2, applied to equation (1.2) (whose USNDS solution is S_0), while considering (1.1) as a nearby problem, corresponding to a perturbation $(-\Delta A_0, -\Delta F_0, -\Delta G_0)$ of the parameters of (1.2). □

THEOREM 4.4. *Assume that the parameters* A, F, G *of* (1.1) *are known exactly, but that* X *cannot be computed exactly. Instead, the exact solution* S_0 *of a nearby problem* (1.2) (*for some parameters* $A+\Delta A_0$, $F+\Delta F_0$, $G+\Delta G_0$) *is available. Assume that* S_0 *is symmetric, nonnegative definite, and stabilizing for the pair* (A, F).

If the perturbation $(\Delta A_0, \Delta F_0, \Delta G_0)$ *is in* P_δ *with*

$$\delta < \delta_2 = \frac{1}{4(M_{S_0}+N_{S_0})}, \tag{4.13}$$

then the matrix S_0 *is the* USNDS *solution to* (1.2). *Moreover,* (1.1) *has a* USNDS *solution* X *such that*

$$\|X-S_0\| \leqq \frac{K_B(S_0)\|S_0\|\delta}{1-2N_{S_0}\delta}, \tag{4.14}$$

where $K_B(S_0)$, M_{S_0}, *and* N_{S_0} *are defined as in* (2.4) *and* (4.1), *except for* S_0 *replacing* X.

Proof. Assume (4.13). By the same argument as in the beginning of the proof of Theorem 4.2 (with S_0 replacing X), the solution S_0 to (1.2) can be shown to be stabilizing for the pair $(A+\Delta A_0, F+\Delta F_0)$, whence it is the USNDS solution to (1.2). Also, since (A, F) is stabilizable, (1.1) has a unique strong solution X.

Let $\Delta X = S_0 - X$. In order to derive (4.14), subtract the two equations

$$(A+\Delta A_0)^T S_0 + S_0(A+\Delta A_0) - S_0(F+\Delta F_0)S_0 + G + \Delta G_0 = 0$$

and

$$A^T(S_0-\Delta X)+(S_0-\Delta X)A-(S_0-\Delta X)F(S_0-\Delta X)+G=0$$

to obtain

$$\Delta X = -\Omega_{S_0}^{-1}(\Delta X F \Delta X) - \Omega_{S_0}^{-1}(\Delta A_0^T S_0 + S_0 \Delta A_0 + \Delta G_0 - S_0 \Delta F_0 S_0).$$

This leads to $\|\Delta X\| \leq \|\Omega_{S_0}^{-1}\| \|F\| \|\Delta X\|^2 + K_B(S_0)\|S_0\|\delta$, which, by an argument similar to that used in Theorem 4.2, implies that

$$\|\Delta X\| \leq \frac{1-\sqrt{1-4N_{S_0}\delta}}{2\|F\|\|\Omega_{S_0}^{-1}\|} = \frac{2K_B(S_0)\|S_0\|\delta}{1+\sqrt{1-4N_{S_0}\delta}}.$$

Finally, noting that $\sqrt{1-x} > 1-x$ for $0<x<1$, we obtain exactly (4.14).

Now, $A-FX$ will be stable if $2\|\Omega_{S_0}^{-1}\|\|F\|\|\Delta X\| < 1$ (from Lemma 3.1), and a fortiori if

$$\frac{2N_{\delta_0}\delta}{1-2N_{S_0}\delta} < 1,$$

using (4.14), and this last inequality does hold for δ satisfying (4.13). □

The last two theorems provide computable bounds for the sensitivity of the USNDS solution to (1.1), involving only an available approximation to this solution. These bounds depend uniformly on the magnitude of data perturbation δ and are valid in a limited range specified by conditions (4.11) or (4.13). This domain of validity shrinks and the error growth factor increases when K_B or $\|\Omega^{-1}\|$ become large, which is perfectly consistent with the idea that very small perturbations can drastically affect the solution when the equation becomes ill-conditioned. In the ill-conditioned case, only very accurate data and a very good model will make S_0 a meaningful approximation of X. In most applications, δ can be identified as a worst-case estimate of the level of inaccuracy corrupting the modeling-solving process. Such an estimate may not be available for the algorithmic part, in which case specific results like [8] can be used as an alternative.

In § 7, this result will be applied to the Newton iterative refinement process. It will be shown that, inside the domain of validity (perturbation-wise), this process is guaranteed to decrease the error after the first step and converge faster than a geometric sequence whose common ratio is proportional to δ. But first, the sensitivity of the ARE condition estimate K_B is analyzed.

5. Bounding the variation of $K_B(X)$. This section focuses on the sensitivity of the condition estimate $K_B(X)$ to perturbations of the parameters of (1.1). Bounds are derived for the error occurring when the exact solution X to (1.1) is replaced by an approximation S, typically solving a nearby problem (1.2). Since any computed solution is only an approximation of X, these results are of practical importance. Indeed, $K_B(X)$ can only be estimated through its counterparts in terms of S and A, F, G or $A+\Delta A$, $F+\Delta F$, $G+\Delta G$, depending on which parameters are known.

The equations (1.1) and (1.2) are assumed to have USNDS solutions X and S, respectively. The following quantities will be compared:

(a) $K_B(X)$ (expressed at X with parameters A, F, G, as defined in (2.4)),

(b) $K_B(S)$ (expressed at S with parameters A, F, G, and with S replacing X in (2.4)),

(c) $\hat{K}_B(S)$, (expressed at S with parameters $A+\Delta A$, $F+\Delta F$, $G+\Delta G$; these parameters and S replacing A, F, G, and X, respectively, in (2.4) and related definitions).

A perturbation analysis generalizing a result in [9] (Lemma 1) is used here. All the needed bounds can be obtained as particular cases of the following theorem on linear operators in **L**.

THEOREM 5.1. *Let Θ and Ω be two operators in* **L**, *and suppose Ω is invertible. Define $\Psi = \Omega^{-1} \circ \Theta$, where "$\circ$" is the composition of operators, and consider linear*

perturbations $\Delta\Omega$, $\Delta\Theta$ *of* Ω, Θ. *By linear, we mean that* $\Delta\Omega \in \mathbf{L}$ *and* $\Delta\Theta \in \mathbf{L}$. *If*

$$\|\Delta\Omega\| < \frac{1}{\|\Omega^{-1}\|}, \tag{5.1}$$

then $\Omega+\Delta\Omega$ *is invertible, and the norm of* $\Psi+\Delta\Psi = (\Omega+\Delta\Omega)^{-1} \circ (\Theta+\Delta\Theta)$ *satisfies*

$$\frac{\|\Psi\| - \|\Omega^{-1}\|\,\|\Delta\Theta\|}{1+\|\Omega^{-1}\|\,\|\Delta\Omega\|} \leqq \|\Psi+\Delta\Psi\| \leqq \frac{\|\Psi\| + \|\Omega^{-1}\|\,\|\Delta\Theta\|}{1-\|\Omega^{-1}\|\,\|\Delta\Omega\|}. \tag{5.2}$$

Proof. That the operator $\Omega+\Delta\Omega$ is invertible when $\Delta\Omega$ satisfies (5.1) follows from [6, Lemma VII.6.1, p. 584].

Let Z be a matrix of unit norm, and V, W, E in $\mathbf{R}^{n\times n}$ be defined as

$$V = \Omega^{-1} \circ \Theta(Z), \qquad W = (\Omega+\Delta\Omega)^{-1} \circ (\Theta+\Delta\Theta)(Z), \quad \text{and} \quad E = V - W.$$

From the definition of V and W, it follows that

$$(\Omega+\Delta\Omega)(W) = (\Theta+\Delta\Theta)(Z) \quad \text{and} \quad \Omega(W+E) = \Theta(Z),$$

and by subtracting,

$$\Omega(E) = \Delta\Omega(W) - \Delta\Theta(Z),$$

whereupon

$$\|E\| \leqq \|\Omega^{-1}\|(\|\Delta\Omega\|\,\|W\| + \|\Delta\Theta\|).$$

Now, $\|W\| - \|E\| \leqq \|V\| \leqq \|W\| + \|E\|$, and therefore

$$\|W\|(1-\|\Omega^{-1}\|\,\|\Delta\Omega\|) \leqq \|V\| + \|\Omega^{-1}\|\,\|\Delta\Theta\| \tag{5.3}$$

and

$$\|V\| - \|\Omega^{-1}\|\,\|\Delta\Theta\| \leqq \|W\|(1+\|\Omega^{-1}\|\,\|\Delta\Omega\|). \tag{5.4}$$

Considering the first inequality (5.3) and recalling that $\|V\| \leqq \|\Psi\|\,\|Z\| = \|\Psi\|$, yields

$$\|W\|(1-\|\Omega^{-1}\|\,\|\Delta\Omega\|) \leqq \|\Psi\| + \|\Omega^{-1}\|\,\|\Delta\Theta\|.$$

Finally, taking the supremum of the left-hand side over matrices Z of unit norm leads to

$$\|\Psi+\Delta\Psi\|(1-\|\Omega^{-1}\|\,\|\Delta\Omega\|) \leqq \|\Psi\| + \|\Omega^{-1}\|\,\|\Delta\Theta\|,$$

which is the second inequality of (5.2). The first one is derived similarly from the inequality (5.4). □

A first application of this theorem is bounding the ratio $(K_B(X)\|X\|)/(K_B(S)\|S\|)$ in terms of $\|X-S\|$.

THEOREM 5.2. *Suppose* (1.1) *and* (1.2) *have* USNDS *solutions* X *and* S, *respectively, and let* $\Delta X = X - S$. *If*

$$\|\Delta X\| < \frac{1}{2\|F\|\,\|\Omega_S^{-1}\|}, \tag{5.5}$$

then $K_B(X)\|X\|$ *and* $K_B(S)\|S\|$ *are related by*

$$\frac{K_B(S)\|S\| - 2M_S\|\Delta X\| - \|F\|\,\|\Omega_S^{-1}\|\,\|\Delta X\|^2}{1+2\|F\|\,\|\Delta X\|\,\|\Omega_S^{-1}\|} \leqq K_B(X)\|X\| \tag{5.6}$$

and

$$K_B(X)\|X\| \leqq \frac{K_B(S)\|S\| + 2M_S\|\Delta X\| + \|F\|\,\|\Omega_S^{-1}\|\,\|\Delta X\|^2}{1-2\|F\|\,\|\Delta X\|\,\|\Omega_S^{-1}\|}. \tag{5.7}$$

Proof. Apply Theorem 5.1 three times to bound the variation of $\|\Omega_X^{-1}\|$, $\|\Theta_X\|$, $\|\Pi_X\|$ successively. For these three applications, Ω and $\Delta\Omega$ are, respectively, set to

$$\Omega = \Omega_S \quad \text{and} \quad \Delta\Omega = \Omega_X - \Omega_S.$$

Since $(\Omega_X - \Omega_S)(Z) = (S-X)FZ + ZF(S-X)$, it follows that

$$\|\Delta\Omega\| = \|\Omega_X - \Omega_S\| \leqq 2\|F\|\,\|X-S\|.$$

This inequality and (5.5) guarantee that the requirement (5.1) of Theorem 5.1 is fulfilled for $\Omega = \Omega_S$. Also, the expressions in the denominators of (5.6)-(5.7) directly follow from the denominator expressions in (5.2). As for the instances of Θ and $\Delta\Theta$, we take successively:

(1) $\Theta(Z) = Z$, $\Delta\Theta = 0$.

(2) $\Theta(Z) = Z^T S + SZ$, $\Delta\Theta(Z) = Z^T(X-S) + (X-S)Z$ $\quad (\|\Delta\Theta\| \leqq 2\|X-S\|)$.

(3) $\Theta(Z) = SZS$, $\Delta\Theta(Z) = XZX - SZS$ $(\|\Delta\Theta\| \leqq 2\|S\|\,\|X-S\| + \|X-S\|^2)$.

In each case, Theorem 5.1 produces lower and upper bounds for $\|\Omega_X^{-1}\|$, $\|\Theta_X\|$, and $\|\Pi_X\|$, respectively. Multiplying these inequalities by $\|G\|$, $\|A\|$, and $\|F\|$, respectively, and adding them together yields the desired bounds (5.6)-(5.7). □

The next theorem relates $K_B(S)$ to $\hat{K}_B(S)$.

THEOREM 5.3. *Let X and S be as in Theorem* 5.2, *and assume the perturbation $(\Delta A, \Delta F, \Delta G)$ is in P_δ (for some $\delta > 0$). Let M_S be defined as in* (4.1) *with S replacing X. If*

$$2M_S\delta < 1, \tag{5.8}$$

then the following inequalities hold:

$$\frac{K_B(S)(1-\delta)}{1+2M_S\delta} \leqq \hat{K}_B(S) \leqq \frac{K_B(S)(1+\delta)}{1-2M_S\delta}. \tag{5.9}$$

Proof. Recall the definition of the operator $\hat{\Omega}_S$:

$$\hat{\Omega}_S(Z) = [A+\Delta A - (F+\Delta F)S]^T Z + Z[A+\Delta A - (F+\Delta F)S].$$

Since $(\hat{\Omega}_S - \Omega_S)(Z) = (\Delta A - \Delta F S)^T Z + Z(\Delta A - \Delta F S)$, we have

$$\|\hat{\Omega}_S - \Omega_S\| \leqq 2(\|A\| + \|F\|\,\|S\|)\delta.$$

Assuming that $2M_S\delta < 1$, Theorem 5.1 can be applied with $\Omega = \Omega_S$, $\Delta\Omega = \hat{\Omega}_S - \Omega_S$ and, successively:

(1) $\Theta = I$, $\Delta\Theta = 0$.

(2) $\Theta(Z) = Z^T S + SZ$, $\Delta\Theta = 0$.

(3) $\Theta(Z) = SZS$, $\Delta\Theta = 0$.

The first instance yields

$$\frac{\|\Omega_S^{-1}\|}{1+2M_S\delta} \leqq \|\hat{\Omega}_S^{-1}\| \leqq \frac{\|\Omega_S^{-1}\|}{1-2M_S\delta}. \tag{5.10}$$

Then combining (5.10) with the output of the other two instances (in a similar way as in the proof of Theorem 5.2) produces

$$\frac{C}{1+2M_S\delta} \leqq \hat{K}_B(S) \leqq \frac{C}{1-2M_S\delta},$$

where

$$C = \|\Omega_S^{-1}\|\,\|G+\Delta G\| + \|\Theta_S\|\,\|A+\Delta A\| + \|\Pi_S\|\,\|F+\Delta F\|.$$

This leads to (5.9), since $(1-\delta)\,\|G\| \leqq \|G+\Delta G\| \leqq (1+\delta)\|G\|$. Similar inequalities hold for $\|A+\Delta A\|$ and $\|F+\Delta F\|$. □

6. Numerical examples. All computations were done in double precision Fortran 77 on a Sun 3/50 with relative machine precision $\varepsilon \approx 1.4 \times 10^{-17}$.

Example 1. Consider the one-dimensional ARE

$$fx^2 - 2ax - g = 0, \quad \text{with } f > 0 \quad \text{and} \quad g > 0. \tag{6.1}$$

This equation always has a unique positive definite solution given by

$$x = \frac{a+\sqrt{a^2+fg}}{f} = \frac{g}{\sqrt{a^2+fg}-a}, \tag{6.2}$$

which is stabilizing, since the closed-loop matrix is

$$a - fx = -\sqrt{a^2+fg} < 0.$$

Define s to be $fx-a$. We then have $x=(a+s)/f$, and also

$$\|\Omega^{-1}\| = \frac{1}{2s}, \qquad \|\Theta\| = \frac{x}{s}, \qquad \|\Pi\| = \frac{x^2}{2s}.$$

This leads to

$$m = \frac{1}{2} + \frac{a+|a|}{2s}, \qquad k_B = 1 + \frac{|a|}{s}, \qquad n = \frac{1}{2}\left(1+\frac{a}{s}\right)\left(1+\frac{|a|}{s}\right).$$

Now consider perturbations Δa, Δf, Δg of the parameters a, f, g, constrained to

$$|\Delta a| \leqq \delta|a|, \qquad |\Delta f| \leqq \delta f, \qquad |\Delta g| \leqq \delta g, \tag{6.3}$$

where δ is a positive number. Since the solution x of (6.1) is an increasing function of a and g (this is immediate when looking at the corresponding partial derivatives), and a decreasing function of f (cf. (6.2)), the maximum variation of x for perturbations constrained by (6.3) will occur at

$$\Delta a = \varepsilon|a|\delta, \qquad \Delta g = \varepsilon g\delta, \qquad \Delta f = -\varepsilon f\delta,$$

where ε is $+1$ or -1. These two sets of perturbations define two perturbed equations whose nonnegative solutions (if any) are denoted by x_+ ($\varepsilon=+1$) and x_- ($\varepsilon=-1$). The largest variation of x for the class of perturbations (6.3) is therefore given by

$$|\Delta x|_{\max} = \max\left(|x-x_+|, |x-x_-|\right). \tag{6.4}$$

In order to test the bound (4.6), a variety of parameter sets (a, f, g) was selected, and, for each set, the parameter δ_0 given by (4.5) and the parameters K_B, M, N were computed. We then considered values of δ (cf. (6.3)) ranging from $10^{-3}\delta_0$ up to $10^2\delta_0$, and computed x_+ and x_- as

$$x_+ = \frac{a+\delta|a|+\sqrt{(a+\delta|a|)^2+(1-\delta^2)fg}}{(1-\delta)f}, \qquad x_- = \frac{a-\delta|a|+\sqrt{(a-\delta|a|)^2+(1-\delta^2)fg}}{(1+\delta)f}.$$

From this data, for a given set a, f, g and for each δ, the maximum variation of x as given by (6.4) could be compared with the value taken by the right-hand expression in the bound (4.6). More precisely, we tested the equivalent of (4.6) in relative terms and thus compared $\Delta x_r := (|\Delta x|_{\max}/x)$ with $B(\delta) := (K_B\delta/1-3(M+N)\delta)$. The outcome of this experiment is displayed in Table 6.1. Each pair of columns corresponds to a certain ratio δ/δ_0. For the largest values of δ, the numbers f and g sometimes become

TABLE 6.1

	$\delta = 10^{-3}\delta_0$		$\delta = 10^{-2}\delta_0$		$\delta = 10^{-1}\delta_0$	
a, f, g	Δx_r	$B(\delta)$	Δx_r	$B(\delta)$	Δx_r	$B(\delta)$
0.1, 2, 2	1.1×10^{-4}	1.1×10^{-4}	1.1×10^{-3}	1.1×10^{-3}	1.1×10^{-2}	1.1×10^{-2}
−0.1, 2, 2	1.1×10^{-4}	1.1×10^{-4}	1.2×10^{-3}	1.2×10^{-3}	1.2×10^{-2}	1.2×10^{-2}
1.0, 2, 2	0.9×10^{-4}	0.9×10^{-4}	0.9×10^{-3}	0.9×10^{-3}	0.9×10^{-2}	0.9×10^{-2}
−1.0, 2, 2	1.8×10^{-4}	1.8×10^{-4}	1.8×10^{-3}	1.8×10^{-3}	1.8×10^{-2}	1.8×10^{-2}
100, 2, 2	0.7×10^{-4}	0.7×10^{-4}	0.7×10^{-3}	0.7×10^{-3}	0.7×10^{-2}	0.7×10^{-2}
−100, 2, 2	4.0×10^{-4}	4.0×10^{-4}	4.0×10^{-3}	4.0×10^{-3}	4.0×10^{-2}	4.1×10^{-2}
10, 2, 20	0.7×10^{-4}	0.7×10^{-4}	0.7×10^{-3}	0.7×10^{-3}	0.7×10^{-2}	0.8×10^{-2}
1, 50, 2	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-2}	1.0×10^{-2}
20, 10, 0.1	0.7×10^{-4}	0.7×10^{-4}	0.7×10^{-3}	0.7×10^{-3}	0.7×10^{-2}	0.7×10^{-2}
	$\delta = \delta_0$		$\delta = 10\delta_0$		$\delta = 100\delta_0$	
a, f, g	Δx_r	$B(\delta)$	Δx_r	$B(\delta)$	Δx_r	$B(\delta)$
0.1, 2, 2	1.1×10^{-1}	1.9×10^{-1}	**	**	**	**
−0.1, 2, 2	1.2×10^{-1}	2.1×10^{-1}	**	**	**	**
1.0, 2, 2	0.9×10^{-1}	1.6×10^{-1}	1.9	**	**	**
−1.0, 2, 2	1.9×10^{-1}	3.1×10^{-1}	1.2	**	**	**
100, 2, 2	0.7×10^{-1}	1.3×10^{-1}	1.1	**	2.8	**
−100, 2, 2	4.9×10^{-1}	6.6×10^{-1}	10^5	**	10^5	**
10, 2, 20	0.8×10^{-1}	1.4×10^{-1}	1.2	**	2.4	**
1, 50, 2	1.0×10^{-1}	1.8×10^{-1}	7.9	**	**	**
20, 10, 0.1	0.7×10^{-1}	1.3×10^{-1}	1.0	**	2.8	**

negative, in which case the perturbed equation no longer falls in the category (6.1). In such cases, double asterisks appear in place of the maximum relative variation of x. We also put double asterisks in the $B(\delta)$ column whenever this bound became a negative number.

These results indicate that the bound (4.6) is tight for relative magnitudes of perturbation δ up to δ_0, with a noticeable difference only when $\delta \approx \delta_0$. Moreover, the threshold δ_0 coincides approximately with the point where the relative variation of X (i.e., $\|\Delta X\|/\|X\|$) starts exceeding 1. Since we are typically interested in keeping this relative variation small, the domain of validity of the bound (4.6), as prescribed by δ_0, does not seem limiting at all.

Example 2. The applications and performance of Theorem 4.4 are illustrated on the following problem of order 3. Consider, for $\lambda = 1.0, 0.1, 0.01$, and 0.001 successively, the triple $(A_\lambda, F_\lambda, I)$ where

$$A_\lambda = \begin{pmatrix} -1.0 & 1.0 & -0.5 \\ 1.0 & 1.0 & 0.5 \\ \lambda & -\lambda & 0.1 \end{pmatrix}; \qquad F_\lambda = \begin{pmatrix} 1.0 & 0.5 & 0.0 \\ 0.5 & 1.0 & 0.0 \\ 0.0 & 0.0 & \lambda \end{pmatrix}. \tag{6.5}$$

For these values of λ, the pairs (A_λ, F_λ) are stabilizable. Let X_λ be the USNDS solution to

$$A_\lambda^T X_\lambda + X_\lambda A_\lambda - X_\lambda F_\lambda X_\lambda + I = 0, \tag{6.6}$$

and denote by S_λ an approximation of X_λ as computed by a Schur solver.

For each value of λ as specified above, estimates $\tilde{K}_B(\lambda)$, $\tilde{M}(\lambda)$, $\tilde{N}(\lambda)$, and $\tilde{\delta}_2(\lambda)$ of $K_B(S_\lambda)$, M_{S_λ}, N_{S_λ}, and δ_2 from (4.13) are obtained as follows (cf. [10, Thm. 2.4]):

(1) Form the closed-loop matrix $K_\lambda = A_\lambda - F_\lambda S_\lambda$.

(2) Compute the solution H_λ to $K_\lambda^T H_\lambda + H_\lambda I_\lambda + I = 0$ and use the identity $\|\Omega_{S_\lambda}^{-1}\| = \|H_\lambda\|$ to estimate $\|\Omega_{S_\lambda}^{-1}\|$.

(3) Compute J_λ solving $K_\lambda^T J_\lambda + J_\lambda K_\lambda + S_\lambda^2 = 0$ and use the identity $\|\Pi_{S_\lambda}\| = \|J_\lambda\|$ to compute $\|\Pi_{S_\lambda}\|$.

(4) Using the inequality $\|\Theta_{S_\lambda}\| \leqq (\|H_\lambda\|\,\|J_\lambda\|)^{1/2}$, estimate $\|\Theta_{S_\lambda}\|$ from above with $(\|H_\lambda\|\,\|J_\lambda\|)^{1/2}$.

(5) Using (2)–(4), (2.4), and (4.1), compute $\tilde{K}_B(\lambda)$, $\tilde{M}(\lambda)$, $\tilde{N}(\lambda)$, and $\tilde{\delta}_2(\lambda)$. Note that $K_B(S_\lambda) \leqq \tilde{K}_B(\lambda)$, $N_{S_\lambda} \leqq \tilde{N}(\lambda)$, and $\tilde{\delta}_2(\lambda) \leqq \delta_2$ of (4.13).

Provided that the computed solution actually solves a problem within $\tilde{\delta}_2(\lambda)$ of the problem (6.6), Theorem 4.4 guarantees that the relative difference between the exact solution X_λ and S_λ is at most

$$u(\lambda) = \frac{\tilde{K}_B(\lambda)\tilde{\delta}_2(\lambda)}{1 - 2\tilde{N}(\lambda)\tilde{\delta}_2(\lambda)}.$$

The magnitudes of $\tilde{\delta}_2(\lambda)$ and $u(\lambda)$ appear in Table 6.2. Note that the safety domain $P_{\tilde{\delta}_2(\lambda)}$ shrinks as λ decreases. This is expected since the pair (A_λ, F_λ) then becomes nearly unstabilizable, thus making the ARE increasingly sensitive. For very ill-conditioned problems, $\tilde{\delta}_2(\lambda)$ will be so small that the stability of the algorithm will no longer guarantee that the ARE solved by S_λ remains in $P_{\tilde{\delta}_2(\lambda)}$. The accuracy guarantee of (4.14) will then be lost.

Finally, Theorem 4.4 can be applied to bound the variation ΔS_λ of S_λ for perturbations of $(A_\lambda, F_\lambda, I)$ in P_δ with

$$\delta < \delta_0(\lambda) = \frac{1}{4 + 8(\tilde{M}(\lambda) + \tilde{N}(\lambda))}. \tag{6.7}$$

We then have

$$\|\Delta S_\lambda\| \leqq \nu(\lambda, \delta) = \frac{\tilde{K}_B(\lambda)\|S_\lambda\|\delta}{1 - 3(\tilde{M}(\lambda) + \tilde{N}(\lambda))\delta}.$$

TABLE 6.2

λ	$\tilde{\delta}_2(\lambda)$	$u(\lambda)$	$\tilde{K}_B(\lambda)$
1.0	1.6×10^{-2}	2.8×10^{-2}	6.3
0.1	1.4×10^{-3}	1.0×10^{-2}	5.4×10^{1}
0.01	1.2×10^{-5}	8.1×10^{-4}	3.2×10^{3}
0.001	1.0×10^{-7}	7.2×10^{-5}	3.0×10^{5}

TABLE 6.3

λ	$\delta\lambda$	δ	$\delta_0(\lambda)$	$\|\Delta S_\lambda\|$	$\nu(\lambda, \delta)$
1.0	1.0×10^{-2}	6.7×10^{-3}	7.7×10^{-3}	5.5×10^{-3}	6.2×10^{-2}
0.1	5.0×10^{-4}	3.3×10^{-4}	6.9×10^{-4}	1.4×10^{-2}	2.2×10^{-2}
0.01	5.0×10^{-6}	3.3×10^{-6}	6.0×10^{-6}	1.0×10^{-2}	1.4×10^{-2}
0.001	5.0×10^{-8}	3.3×10^{-8}	4.9×10^{-8}	1.0×10^{-2}	1.4×10^{-2}

As an illustration, consider the perturbed problems obtained by replacing λ in (6.5) by $\lambda+\delta\lambda$ with $\delta\lambda$ chosen so that (6.7) is satisfied for the corresponding $\delta = (|\delta\lambda|/\min(\|A_\lambda\|, \|F_\lambda\|)$. The USNDS solution $S_\lambda+\Delta S_\lambda$ to the perturbed problem is computed, and $\|\Delta S_\lambda\|$ is compared to $\nu(\lambda, \delta)$ in Table 6.3. The results in Table 6.3 indicate that the bound (4.6) is realistic within the region $P_{\delta_0(\lambda)}$.

7. Application to the Newton refinement scheme. To conclude this paper, the error bounds derived in § 4 are applied to the Newton method for refining the solution of an ARE. Specifically, a region and a speed of convergence are computed for this scheme in terms of the magnitude of data perturbation δ.

First, recall the fundamental result in [9] concerning the convergence properties of the Newton refinement scheme.

THEOREM 7.1. *Assume* (1.1) *has a* USNDS *solution* X, *and let* S_0 *be an "initial guess" for* X. *Furthermore, let* S_1 *be the refined solution after one iteration of the Newton refinement scheme started at* S_0, *i.e., the solution of*

$$(A-FS_0)^TS_1+S_1(A-FS_0) = -S_0FS_0-G.$$

If

$$\|X-S_0\| \leqq r = \frac{1}{3\|F\|\,\|\Omega_X^{-1}\|}, \tag{7.1}$$

then

$$\|X-S_1\| \leqq [\|F\|\,\|\Omega_{S_0}^{-1}\|\,\|X-S_0\|]\|X-S_0\|. \tag{7.2}$$

Note that the Newton iterations are always converging to the exact solution X provided that the initial guess S_0 is stabilizing for the pair (A, F) (see [11]). Nevertheless, without condition (7.1), the error may be drastically increased during the first iteration before decreasing to zero. When (7.1) holds, however, the growth factor $\rho_0 = \|F\|\,\|\Omega_{S_0}^{-1}\|\,\|X-S_0\|$ is less than 1, and therefore S_1 is guaranteed to be a better estimate of X than S_0. Note that

$$\rho_0 \leqq k = \frac{\|F\|\,\|\Omega_X^{-1}\|\,\|X-S_0\|}{1-2\|F\|\,\|\Omega_X^{-1}\|\,\|X-S_0\|} \ (<1). \tag{7.3}$$

More generally, the errors after i and $i+1$ steps are related by

$$\|X-S_{i+1}\| \leqq \rho_i\|X-S_i\|,$$

where

$$\rho_i \leqq \frac{\|F\|\,\|\Omega_X^{-1}\|\,\|X-S_i\|}{1-2\|F\|\,\|\Omega_X^{-1}\|\,\|X-S_i\|} < 1.$$

Since the error $\|X-S_i\|$ is decreasing with i, all the ρ_i are less than k, and thus the error decreases to zero at least as fast as the sequence $\{k^n\}_{n\geqq 1}$.

In the remainder of this section, the bound of Theorem 4.2 is used to express the result of Theorem 7.1 directly in terms of δ.

THEOREM 7.2. *With the notation of Theorems* 4.2 *and* 7.1, *and if*

$$\delta < \frac{1}{4+8(M_X+N_X)}, \tag{7.4}$$

then

$$\|X - S_1\| \leqq R(\delta)\|X - S_0\|,$$

where

$$R(\delta) = \frac{N_X \delta}{1 - (3M_X + 5N_X)} \quad \textit{and} \quad R(\delta) < 1. \tag{7.5}$$

Proof. The assumption (7.4) ensures that Theorem 4.2 applies and (4.6) holds. But (7.4) also implies that $3N_X\delta < 1 - 3(M_X + N_X)\delta$, or equivalently that

$$\frac{K_B(X)\|X\|\delta}{1 - 3(M_X + N_X)\delta} \leqq \frac{1}{3\|F\|\,\|\Omega_X^{-1}\|}.$$

This last result, combined with (4.6), ensures that condition (7.1) of Theorem 7.1 is satisfied, and thus

$$\|X - S_1\| \leqq \frac{\|F\|\,\|\Omega_X^{-1}\|\,\|X - S_0\|}{1 - 2\|F\|\,\|\Omega_X^{-1}\|\,\|X - S_0\|}\|X - S_0\|.$$

But using (4.6) again yields

$$\frac{\|F\|\,\|\Omega_X^{-1}\|\,\|X - S_0\|}{1 - 2\|F\|\,\|\Omega_X^{-1}\|\,\|X - S_0\|} \leqq \frac{\dfrac{N_X\delta}{1 - 3(M_X + N_X)\delta}}{1 - 2\left(\dfrac{N_X\delta}{1 - 3(M_X + N_X)\delta}\right)}$$

$$= \frac{N_X\delta}{1 - (3M_X + 5N_X)\delta} = R(\delta).$$

Finally, $R(\delta) < 1$ follows trivially from (7.4). □

Note that the bounds of Theorems 4.3 and 4.4 could have been used instead in order to produce counterparts of (7.4)-(7.5) depending only on computable quantities. Theorem 7.2 has an important implication. That is, in the entire range of perturbations for which Theorem 4.2 applies, the Newton refinement iterations started at the computed solution S_0 are guaranteed not only to converge to X, but also to decrease the error after the first step and finally to converge faster than a geometric sequence of common ratio $R(\delta) < 1$. Therefore, the threshold δ_0 in (4.5) not only makes the error magnitude $\|X - S\|$ tractable in terms of δ, but it also ensures the fast convergence of the Newton refinement process when started at the solution S of any problem (1.2) with perturbations in P_{δ_0}. In other words, δ_0 defines a safety region (perturbation-wise), within which accurate approximations of X are guaranteed when the ARE is solved with a stable algorithm followed by a Newton iterative refinement.

Conclusions. The computation of the USNDS solution to the symmetric ARE (1.1) was shown to be a well-posed problem, in the sense that if such a solution exists for a particular set of parameters (A, F, G), then it exists in an open neighborhood of (A, F, G) and depends continuously on the parameters in this neighborhood. Computable estimates for the sensitivity of the USNDS solution to parameter perturbation were also derived. They indicate what accuracy can at least be expected for the computed solution provided that the magnitude of the parameter variations does not exceed some explicit threshold. Note that "parameter variations" encompass the error occurring at the identification level, as well as the perturbation introduced during the computation.

REFERENCES

[1] W. F. ARNOLD AND A. J. LAUB, *Generalized eigenproblem algorithms and software for algebraic Riccati equations*, Proc. IEEE, 72 (1984), pp. 1746–1754.

[2] R. BYERS, *Numerical condition of the algebraic Riccati equation*, in Proc. Summer Research Conference, AMS Contemporary Math., vol. 47, Amer. Math. Soc., Providence, RI, 1984, pp. 35–49.

[3] F. CALLIER AND J. WILLEMS, *Criterion for the convergence of the solution of the Riccati differential equation*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1232–1242.

[4] S. W. CHAN, G. C. GOODWIN, AND K. S. SIN, *Convergence properties of the Riccati difference equation in optimal filtering of nonstabilizable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 110–118.

[5] D. F. DELCHAMPS, *A note on the analyticity of the Riccati metric*, in Algebraic and Geometric Methods in Linear Systems Theory, Lecture Notes in Applied Mathematics 18, Amer. Math. Soc., Providence, RI, 1980, pp. 37–41.

[6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, vol. 1, Interscience Publishers Inc., New York, 1967.

[7] G. HEWER AND C. KENNEY, *The sensitivity of the stable Lyapunov equation*, SIAM J. Control Optim., 26 (1988), pp. 321–344.

[8] C. KENNEY, A. J. LAUB, AND M. WETTE, *A stability-enhancing scaling procedure for Schur–Riccati solvers*, System Control Lett., 12 (1989), pp. 241–250.

[9] ———, *Error bounds for Newton refinement of solutions to algebraic Riccati equations*, Math. Contr. Sig. Syst., 3 (1990), pp. 211–224.

[10] C. KENNEY AND G. HEWER, *The sensitivity of the algebraic and differential Riccati equations*, SIAM J. Control Optim., 28 (1990), pp. 50–69.

[11] D. L. KLEINMAN, *On an iterative technique for Riccati equation computations*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 114–115.

[12] V. KUCERA, *A contribution to matrix quadratic equations*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 344–347.

[13] A. J. LAUB, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 913–921.

[14] M.-A. POUBELLE, I. R. PETERSEN, M. R. GEVERS, AND R. R. BITMEAD, *A miscellany of results on an equation of Count J. F. Riccati*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 651–654.

[15] J. RICE, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.

[16] H. SAGAN, *Advanced Calculus*, Houghton Mifflin, Boston, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1481-1490, November 1990

012

LYAPUNOV FUNCTIONS AND ALMOST SURE EXPONENTIAL STABILITY OF STOCHASTIC DIFFERENTIAL EQUATIONS BASED ON SEMIMARTINGALES WITH SPATIAL PARAMETERS*

XUERONG MAO†

This paper is dedicated to Professor K. D. Elworthy and Mrs. S. M. Elworthy on the occasion of their 25th wedding anniversary.

Abstract. The objective of this paper is to use the Lyapunov function to study the almost sure exponential stability of the stochastic differential equation

$$\varphi_t = x + \int_0^t F(\varphi_s, ds),$$

where $F(x, t)$ is a continuous C-semimartingale with spatial parameter x. This equation includes many important stochastic systems, for example, the classical Itô equation. More importantly, the result can be employed to study the bound of the Lyapunov exponent of stochastic flows.

Key words. Lyapunov function, almost sure exponential stability, stochastic differential equation, semimartingale, exponential martingale inequality

AMS(MOS) subject classifications. 60H, 93D

1. Introduction. Numerous problems in science and engineering lead to the study of the exponential stability of stochastic systems. Has'minskii [8] gave a necessary and sufficient criterion for almost sure exponential stability of the linear Itô equation which opened a new chapter in stochastic stability theory. In fact, there exists an extensive literature in this area, in particular, we mention Arnold [1], Arnold and Kliemann [2], Arnold, Oeljeklaus, and Pardoux [3], Caverhill [4], Chappell [5], Crauel [6], and Curtain [7]. However, it seems there is almost no work being done by using the Lyapunov function to study the almost sure exponential stability of stochastic systems, although a few papers exist, for instance, Ladde [12], which employ the Lyapunov function to deal with the moment exponential stability.

The objective of this paper is to study the almost sure exponential stability of the stochastic differential equation

$$\varphi_t = x + \int_0^t F(\varphi_s, ds) \tag{1.1}$$

via the Lyapunov function, where F is a continuous C-semimartingale with spatial parameter x. We would like to mention that (1.1) includes many important stochastic systems. For instance, if we let

$$F(x, t) = \int_0^t f(x, s)\, dN_s + \int_0^t b(x, s)\, dA_s, \qquad t \geqq 0,$$

where N is an m-dimensional continuous local martingale and A a continuous non-decreasing adapted process, then (1.1) reduces to the familiar stochastic differential

* Received by the editors June 26, 1989; accepted for publication (in revised form) January 14, 1990.
† Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom. This research was supported by grant GR/F51241 of the Science and Engineering Research Council.

equation with respect to semimartingales

$$\varphi_t = x + \int_0^t f(\varphi_s, s)\, dN_s + \int_0^t b(\varphi_s, s)\, dA_s, \qquad t \geqq 0. \tag{1.2}$$

If we let

$$F(x, t) = \int_0^t f(x, s)\, dW_s + \int_0^t b(x, s)\, ds, \qquad t \geqq 0,$$

where W is an m-dimensional Wiener process, then (1.1) reduces to the classical Itô equation

$$\varphi_t = x + \int_0^t f(\varphi_s, s)\, dW_s + \int_0^t b(\varphi_s, s)\, ds, \qquad t \geqq 0. \tag{1.3}$$

Hence, as a direct application, we get a sufficient criterion for almost sure exponential stability of (1.2) and (1.3). More importantly, our result can also be employed to study the Lyapunov exponent of stochastic flows. In fact, given a forward stochastic flow of homomorphisms $\varphi_{s,t}(x)$, $0 \leqq s \leqq t < \infty$, $x \in R^n$ under some suitable conditions, we can find a semimartingale $F(x, t)$ with spatial parameter x such that the flow is governed by Itô's stochastic differential equation based on $F(x, t)$, i.e.,

$$\varphi_{s,t}(x) = x + \int_s^t F(\varphi_{s,r}(x), dr) \quad \text{a.s. on } 0 \leqq s \leqq t < \infty, \quad x \in R^n. \tag{1.4}$$

Therefore, we can use the Lyapunov function to estimate the Lyapunov exponent of the stochastic flow. In addition, some examples are worked out to illustrate our results.

2. Main results. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geqq 0}, P)$ be a complete probability space with the right continuous filtration $\{\mathcal{F}_t\}$ containing all P-null sets of $\mathcal{F}$. Let $F(x, t) = (F^1(x, t), \cdots, F^n(x, t))^T$, $(x, t) \in R^n \times R_+$, be a continuous C-semimartingale with spatial parameters, i.e., $F(x, t)$ is a continuous semimartingale for any $x \in R^n$ and is continuous in x for any t almost surely. Let $F^i(x, t) = M^i(x, t) + B^i(x, t)$ be the decomposition such that $M^i(x, t)$ is a continuous local martingale and $B^i(x, t)$ is a continuous process of bounded variation. Set

$$A^{ij}(x, y, t) = \langle M^i(x, t), M^j(y, t) \rangle, \qquad 1 \leqq i, j \leqq n.$$

Then there exists a continuous strictly increasing process A_t with $A_0 = 0$ such that all $A^{ij}(x, y, t)$ and $B^i(x, t)$ are absolutely continuous with respect to A_t almost surely for any $x, y \in R^n$. Therefore, there exist predictable processes $a^{ij}(x, y, t)$ and $b^i(x, t)$ with parameters x, y such that

$$A^{ij}(x, y, t) = \int_0^t a^{ij}(x, y, s)\, dA_s,$$

$$B^i(x, t) = \int_0^t b^i(x, s)\, dA_s.$$

Set $a(x, y, t) = (a^{ij}(x, y, t))_{n \times n}$ and $b(x, t) = (b^1(x, t), \cdots, b^n(x, t))^T$. The triple $(a(x, y, t), b(x, t), A_t)$ is called the characteristic of $F(x, t)$ (cf. Kunita [9]).

It is well known (cf. Kunita [9]) that if φ_t is an n-dimensional predictable process satisfying

$$\int_0^t a^{ii}(\varphi_s, \varphi_s, s)\, dA_s < \infty \qquad \text{a.s.},$$

$$\int_0^t |b^i(\varphi_s, s)|\, dA_s < \infty \qquad \text{a.s.}$$

for all $t \in R_+$, $i = 1, \cdots, n$, then Itô's stochastic integral of φ_t based on the kernel $F(x, dt)$ of the form

$$\int_0^t F(\varphi_s, ds)$$

is well defined. Kunita [9] also discussed the existence and uniqueness of the solution to the stochastic differential equation based on $F(x, t)$ of the type

$$\varphi_t = x + \int_0^t F(\varphi_s, ds). \tag{2.1}$$

Throughout this paper we assume that the equation satisfies the condition of the existence and uniqueness of the solution.

Let $C^2(R^n)$ denote the family of all functions $V: R^n \to R$ with continuous second partial derivatives. Let $\mathcal{P}(R^n \times R_+)$ be the family of all predictable processes $g(x, t)$, $t \geqq 0$ with parameter $x \in R^n$. Define the operator $L: C^2(R^n) \to \mathcal{P}(R^n \times R_+)$ by

$$LV(x) := \sum_{i=1}^n \frac{\partial}{\partial x_i} V(x) b^i(x, t) + \frac{1}{2} \sum_{i,j=1}^n \frac{\partial^2}{\partial x_i \, \partial x_j} V(x) a^{ij}(x, x, t). \tag{2.2}$$

THEOREM 2.1. *Assume there exist a function $V \in C^2(R^n)$, a polynomial $\mu(t)$ $(t \geqq 0)$ with positive coefficients, and positive constants p, λ, α, β, σ such that*

(1) $|x|^p \leqq V(x)$, $x \in R^n$;

(2) $LV(x) \leqq -\lambda V(x) + \mu(t)\, e^{-\lambda \alpha t}$, $(x, t) \in R^n \times R_+$;

(3) $$\sum_{i,j=1}^n \frac{\partial}{\partial x_i} V(x) \frac{\partial}{\partial x_j} V(x) a^{ij}(x, x, t) \leqq \mu(t)\, e^{-\lambda \alpha t} V(x), \qquad (x, t) \in R^n \times R_+;$$

(4) $A_t \leqq \alpha t + \beta$ *almost surely for all* $t \geqq 0$ *and* $\liminf_{t \to \infty} A_t / t \geqq \sigma$ *almost surely. Then the solution of* (2.1) *satisfies*

$$\limsup_{t \to \infty} \frac{1}{t} \log |\varphi_t| \leqq -\lambda \sigma / \rho \quad \text{a.s.} \tag{2.3}$$

Proof. By Itô's formula and assumption (2),

$$\begin{aligned} e^{\lambda A_t} V(\varphi_t) &= V(x) + \int_0^t e^{\lambda A_s} [\lambda V(\varphi_s) + LV(\varphi_s)]\, dA_s \\ &\quad + \int_0^t e^{\lambda A_s} \sum_{i=1}^n \frac{\partial}{\partial x_i} V(\varphi_s) M^i(\varphi_s, ds) \\ &\leqq V(x) + \int_0^t e^{\lambda A_s} \mu(s)\, e^{-\lambda \alpha s}\, dA_s + \int_0^t e^{\lambda A_s} \sum_{i=1}^n \frac{\partial}{\partial x_i} V(\varphi_s) M^i(\varphi_s, ds). \end{aligned} \tag{2.4}$$

By condition (4) we have

$$e^{\lambda A_s - \lambda \alpha s} \leqq e^{\lambda \beta}, \qquad s \geqq 0. \tag{2.5}$$

Therefore,

$$e^{\lambda A_t} V(\varphi_t) \leqq V(x) + e^{\lambda \beta} \int_0^t \mu(s)\, dA_s + \int_0^t e^{\lambda A_s} \sum_{i=1}^n \frac{\partial}{\partial x_i} V(\varphi_s) M^i(\varphi_s, ds). \tag{2.6}$$

Thanks to the exponential martingale inequality, we have

$$P\Bigg[\omega:\sup_{0\leqq t\leqq\tau}\Bigg\{\int_0^t e^{\lambda A_s}\sum_{i=1}^n\frac{\partial}{\partial x_i}V(\varphi_s)M^i(\varphi_s,ds)$$

$$-\frac{\gamma}{2}\int_0^t e^{2\lambda A_s}\sum_{i,j=1}^n\frac{\partial}{\partial x_i}V(\varphi_s)\frac{\partial}{\partial x_j}V(\varphi_s)a^{ij}(\varphi_s,\varphi_s,s)\,dA_s\Bigg\}>\delta\Bigg]\leqq e^{-\gamma\delta} \tag{2.7}$$

for any positive constants γ, δ, and τ. Let $\theta>1$ be arbitrary and take

$$\gamma=\theta^{-(d+1)k},\quad \delta=\theta^{(d+1)k+1}\log k,\quad \tau=\theta^k\quad (k=1,2,\cdots)$$

in (2.7), where d is chosen as d greater then or equal to the degree of $\mu(\cdot)$. Applying the Borel-Cantelli lemma we deduce that there exists an integer $k_0(\omega)$ for almost sure $\omega\in\Omega$ such that

$$\int_0^t e^{\lambda A_s}\sum_{i=1}^n\frac{\partial}{\partial x_i}V(\varphi_s)M^i(\varphi_s,ds)$$

$$\leqq\theta^{(d+1)k+1}\log k+\frac{1}{2}\theta^{-(d+1)k}\int_0^t e^{2\lambda A_s}\sum_{i,j=1}^n\frac{\partial}{\partial x_i}V(\varphi_s)\frac{\partial}{\partial x_j}V(\varphi_s)a^{ij}(\varphi_s,\varphi_s,s)\,dA_s$$

for all $0\leqq t\leqq\theta^k$, $k\geqq k_0$. By assumption (3) and inequality (2.5),

$$\sum_{i,j=1}^n\frac{\partial}{\partial x_i}V(\varphi_s)\frac{\partial}{\partial x_j}V(\varphi_s)a^{ij}(\varphi_s,\varphi_s,s)\leqq\mu\,e^{\lambda\beta-\lambda A_s}V(\varphi_s).$$

It then follows that

$$\int_0^t e^{\lambda A_s}\sum_{i=1}^n\frac{\partial}{\partial x_i}V(\varphi_s)M^i(\varphi_s,ds)\leqq\theta^{(d+1)k+1}\log k$$

$$+\frac{1}{2}\theta^{-(d+1)k}\,e^{\lambda\beta}\int_0^t\mu(s)\,e^{\lambda A_s}V(\varphi_s)\,dA_s$$

for all $0\leqq t\leqq\theta^k$, $k\geqq k_0$ almost surely. Putting this into (2.6) we arrive at

$$e^{\lambda A_t}V(\varphi_t)\leqq V(x)+\theta^{(d+1)k+1}\log k+e^{\lambda\beta}\int_0^t\mu(s)\,dA_s$$

$$+\frac{1}{2}\theta^{-(d+1)k}\,e^{\lambda\beta}\int_0^t\mu(s)\,e^{\lambda A_s}V(\varphi_s)\,dA_s \tag{2.8}$$

for all $0\leqq t\leqq\theta^k$, $k\geqq k_0$ almost surely. In view of Theorem 1 of Mao [10] we get

$$e^{\lambda A_t}V(\varphi_t)\leqq[V(x)+\theta^{(d+1)k+1}\log k+e^{\lambda\beta}\mu(\theta^k)A_0k]\exp\{\tfrac{1}{2}\theta^{-(d+1)k}\,e^{\lambda\beta}\mu(\theta^k)A_0k\}$$

$$\leqq C(1+\theta^{(d+1)k+1}\log k),\qquad 0\leqq t\leqq\theta^k,\quad k\geqq k_0,\quad\text{a.s.}, \tag{2.9}$$

where C is a positive constant independent of k. Consequently,

$$\frac{e^{tA_t}V(\varphi_t)}{t^{d+1}\log\log t}\leqq\frac{C(1+\theta^{(d+1)k+1}\log k)}{\theta^{(d+1)(k-1)}(\log(k-1)+\log\log\theta)},\qquad\theta^{k-1}\leqq t\leqq\theta^k,\quad k\geqq k_0\quad\text{a.s.},$$

which implies

$$\limsup_{t\to\infty}\frac{e^{tA_t}V(\varphi_t)}{t^{d+1}\log\log t}\leqq C\theta^{d+2}\quad\text{a.s.}$$

Letting θ tend to 1 we have

$$\limsup_{t\to\infty}\frac{e^{tA_t}V(\varphi_t)}{t^{d+1}\log\log t}\leqq C\quad\text{a.s.} \tag{2.10}$$

Finally,

$$\begin{aligned}\limsup_{t\to\infty}\frac{1}{t}\log|\varphi_t| &\leqq \limsup_{t\to\infty}\frac{1}{pt}\log V(\varphi_t)\\ &\leqq \limsup_{t\to\infty}\frac{1}{pt}\log\left[e^{-\lambda A_t}t^{d+1}\log\log t\frac{e^{tA_t}V(\varphi_t)}{t^{d+1}\log\log t}\right]\\ &= -\frac{\lambda}{p}\liminf_{t\to\infty}\frac{A_t}{t}\leqq -\lambda\sigma/p \quad \text{a.s.},\end{aligned}$$

which is the desired result and the proof is complete.

THEOREM 2.2. *Assume there exist a function* $V\in C^2(R^n)$ *and positive constants* p, λ, μ, ρ, α, β, σ *such that*

(1) $|x|^p\leqq V(x)$, $x\in R^n$;

(2) $LV(x)\leq -\lambda V(x)+\mu e^{-\lambda\alpha t+\rho t}$, $(x,t)\in R^n\times R_+$;

(3) $\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}V(x)\frac{\partial}{\partial x_j}V(x)a^{ij}(x,x,t)\leqq\mu e^{-\lambda\alpha t+\rho t}V(x)$, $\quad (x,t)\in R^n\times R_+$;

(4) $A_t\leqq\alpha t+\beta$ *almost surely for all* $t\geqq 0$ *and* $\liminf_{t\to\infty}A_t/t\geqq\sigma$ *almost surely. Then the solution of* (2.1) *satisfies*

$$\limsup_{t\to\infty}\frac{1}{t}\log|\varphi_t|\leqq -(\lambda\sigma-\rho)/p \quad a.s. \tag{2.11}$$

Proof. By Itô's formula and conditions (2) and (4),

$$e^{\lambda A_t}V(\varphi_t)\leqq V(x)+\int_0^t\mu e^{\lambda\beta+\rho s}\,dA_s+\int_0^t e^{\lambda A_s}\sum_{i=1}^{n}\frac{\partial}{\partial x_i}V(\varphi_s)M^i(\varphi_s,ds). \tag{2.12}$$

Let $\theta>1$ arbitrarily. Taking

$$\gamma=k^{-1}e^{-\rho k},\quad \delta=k\theta e^{\rho k}\log k,\quad \tau=k\quad (k=1,2,\cdots)$$

in (2.7) and applying the Borel-Cantelli lemma, we deduce that there exists an integer $k_0(\omega)$ for almost sure $\omega\in\Omega$ such that

$$\int_0^t e^{\lambda A_s}\sum_{i=1}^{n}\frac{\partial}{\partial x_i}V(\varphi_s)M^i(\varphi_s,ds)\leqq k\theta e^{\rho k}\log k+\frac{1}{2}e^{\lambda\beta}k^{-1}e^{-\rho k}\int_0^t e^{\rho s}e^{\lambda A_s}V(\varphi_s)\,dA_s$$

for all $0\leqq t\leqq k$, $k\geqq k_0$ almost surely. Hence we get from (2.12) that

$$\begin{aligned}e^{\lambda A_t}V(\varphi_t)\leqq V(x)&+k\theta e^{\rho k}\log k+\int_0^t\mu e^{\lambda\beta+\rho s}\,dA_s\\ &+\frac{1}{2}e^{\lambda\beta}k^{-1}e^{-\rho k}\int_0^t e^{\rho s}e^{\lambda A_s}V(\varphi_s)\,dA_s\end{aligned} \tag{2.13}$$

for all $0\leqq t\leqq k$, $k\geqq k_0$ almost surely, which implies

$$e^{\lambda A_t}V(\varphi_t)\leqq C(1+k\theta e^{\rho k}\log k),\qquad 0\leqq t\leqq k,\quad k\geqq k_0 \quad \text{a.s.}, \tag{2.14}$$

where C is a positive constant. Consequently,

$$\frac{e^{\lambda A_t}V(\varphi_t)}{e^{\rho t}t\log t}\leqq\frac{C(1+k\theta e^{\rho k}\log k)}{e^{\rho(k-1)}(k-1)\log(k-1)},\qquad k-1\leqq t\leqq k,\quad k\geqq k_0 \quad \text{a.s.},$$

which implies immediately that

$$\limsup_{t\to\infty}\frac{e^{\lambda A_t}V(\varphi_t)}{e^{\rho t}t\log t}\leqq Ce^{\rho} \quad \text{a.s.}, \tag{2.15}$$

and then it follows that

$$\limsup_{t\to\infty}\frac{1}{t}\log|\varphi_t|\leqq-(\lambda\sigma-\rho)/p \quad \text{a.s.}$$

The proof is complete.

The following theorem combines Theorems 2.1 and 2.2.

THEOREM 2.3. *Assume there exist a function* $V\in C^2(R^n)$, *a polynomial* $\mu(t)$ $(t\geqq 0)$ *with positive coefficients, positive constants* p, λ, α, β, σ, *and a nonnegative constant* ρ *such that*

(1) $|x|^p\leqq V(x)$, $x\in R^n$;

(2) $LV(x)\leqq-\lambda V(x)+\mu(t)\,e^{-\lambda\alpha t+\rho t}$, $(x,t)\in R^n\times R_+$;

(3) $$\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}V(x)\frac{\partial}{\partial x_j}V(x)a^{ij}(x,x,t)\leqq\mu(t)\,e^{-\lambda\alpha t+\rho t}V(x), \qquad (x,t)\in R^n\times R_+;$$

(4) $A_t\leqq\alpha t+\beta$ *almost surely for all* $t\geqq 0$ *and* $\liminf_{t\to\infty}A_t/t\geqq\sigma$ *almost surely. Then the solution of* (2.1) *satisfies*

$$\limsup_{t\to\infty}\frac{1}{t}\log|\varphi_t|\leqq-(\lambda\sigma-\rho)/p \quad \text{a.s.}$$

Proof. For any $\varepsilon>0$, there exists a constant C such that

$$\mu(t)\,e^{-\lambda\alpha t+\rho t}\leqq C\,e^{-\lambda\alpha t+(\rho+\varepsilon)t}.$$

Hence, by Theorem 2.2,

$$\limsup_{t\to\infty}\frac{1}{t}\log|\varphi_t|\leqq-(\lambda\sigma-\rho-\varepsilon)/p.$$

The conclusion follows since ε is arbitrary.

3. Important corollaries. First we have the following useful corollary, which follows directly from Theorem 2.3.

COROLLARY 3.1. *Assume there exist a positive defined* $n\times n$ *matrix* Q, *a polynomial* $\mu(t)(t\geqq 0)$ *with positive coefficients, positive constants* λ, α, β, σ, *and a nonnegative constant* ρ *such that*

(1) $x^T(Q+Q^T)b(x,t)\leqq-\lambda x^TQx+\mu(t)\,e^{-\lambda\alpha t+\rho t}$, $(x,t)\in R^n\times R_+$;

(2) $\|a(x,x,t)\|\leqq\mu(t)\,e^{-\lambda\alpha t+\rho t}$, $(x,t)\in R^n\times R_+$;

(3) $A_t\leqq\alpha t+\beta$ *almost surely for all* $t\geqq 0$ *and* $\liminf_{t\to\infty}A_t/t\geqq\sigma$ *almost surely. Then the solution of* (2.1) *satisfies*

$$\limsup_{t\to\infty}\frac{1}{t}\log|\varphi_t|\leqq-(\lambda\sigma-\rho)/2 \quad \text{a.s.}$$

We now let $N_t=(N_t^1,\cdots,N_t^m)^T$, $t\geqq 0$, be an m-dimensional continuous martingale such that $N_0=0$ and

$$\langle N^i,N^j\rangle_t=\int_0^t K^{ij}(s)\,dA_s, \qquad t\geqq 0,\quad 1\leqq i,j\leqq m,$$

where K^{ij}, $1\leqq i,j\leqq m$ are all predictable processes. Let $f(x,t)=(f^{ij}(x,t))_{n\times m}$, $t\geqq 0$ be a predictable matrix for each $x\in R^n$. Consider the stochastic differential equation

$$\varphi_t=x+\int_0^t f(\varphi_s,s)\,dN_s+\int_0^t b(\varphi_s,s)\,dA_s, \qquad t\geqq 0. \tag{3.1}$$

Assume the equation satisfies the condition of existence and uniqueness of the solution (cf. Mao and Wu [11]). Note that (3.1) is equivalent to (2.1) if we set

$$F(x, t) = \int_0^t f(x, s)\, dN_s + \int_0^t b(x, s)\, dA_s, \qquad t \geqq 0.$$

Therefore, we have the following corollary.

COROLLARY 3.2. *Assume there exist a function $V \in C^2(R^n)$, a polynomial $\mu(t)(t \geqq 0)$ with positive coefficients, positive constants p, λ, α, β, σ, and a nonnegative constant ρ such that*

(1) $|x|^p \leqq V(x)$, $x \in R^n$;

(2) $$\sum_{i=1}^n \frac{\partial}{\partial x_i} V(x) b^i(x, t) + \frac{1}{2} \sum_{i,j=1}^n \sum_{l,k=1}^m \frac{\partial^2}{\partial x_i\, \partial x_j} V(x) f^{il}(x, t) K^{lk}(t) f^{jk}(x, t)$$
$$\leqq -\lambda V(x) + \mu(t)\, e^{-\lambda\alpha t + \rho t}, \qquad (x, t) \in R^n \times R_+;$$

(3) $$\sum_{i,j=1}^n \sum_{l,k=1}^m \frac{\partial}{\partial x_i} V(x) \frac{\partial}{\partial x_j} V(x) f^{il}(x, t) K^{lk}(t) f^{jk}(x, t)$$
$$\leqq \mu(t)\, e^{-\lambda\alpha t + \rho t} V(x), \qquad (x, t) \in R^n \times R_+;$$

(4) *$A_t \leqq \alpha t + \beta$ almost surely for all $t \geqq 0$ and $\liminf_{t\to\infty} A_t / t \geqq \sigma$ almost surely. Then the solution of* (3.1) *satisfies*

$$\limsup_{t\to\infty} \frac{1}{t} \log |\varphi_t| \leqq -(\lambda\sigma - \rho)/p \quad a.s.$$

More specially, we consider Itô's equation

$$\varphi_t = x + \int_0^t f(\varphi_s, s)\, dW_s + \int_0^t b(\varphi_s, s)\, ds, \qquad t \geqq 0, \tag{3.2}$$

where W is an m-dimensional Wiener process. We then have Corollary 3.3.

COROLLARY 3.3. *Assume there exist a function $V \in C^2(R^n)$, a polynomial $\mu(t)(t \geqq 0)$ with positive coefficients, and constants $p > 0$, $\lambda > 0$, and $\rho \geqq 0$ such that*

(1) $|x|^p \leqq V(x)$, $x \in R^n$;

(2) $$\sum_{i=1}^n \frac{\partial}{\partial x_i} V(x) b^i(x, t) + \frac{1}{2} \sum_{i,j=1}^n \sum_{k=1}^m \frac{\partial^2}{\partial x_i\, \partial x_j} V(x) f^{ik}(x, t) f^{jk}(x, t)$$
$$\leqq -\lambda V(x) + \mu(t)\, e^{-\lambda t + \rho t}, \qquad (x, t) \in R^n \times R_+;$$

(3) $$\sum_{i,j=1}^n \sum_{k=1}^m \frac{\partial}{\partial x_i} V(x) \frac{\partial}{\partial x_j} V(x) f^{ik}(x, t) f^{jk}(x, t)$$
$$\leqq \mu(t)\, e^{-\lambda t + \rho t} V(x), \qquad (x, t) \in R^n \times R_+.$$

Then the solution of (3.2) *satisfies*

$$\limsup_{t\to\infty} \frac{1}{t} \log |\varphi_t| \leqq -(\lambda - \rho)/p \quad a.s.$$

4. The bound for Lyapunov exponents of stochastic flows. In this section we will apply our results to study the bound for Lyapunov exponents of stochastic flows. For the readers' convenience, let us first give the definition of the stochastic flow of homomorphisms and the Brownian flow (cf. Kunita [9]).

Let $\varphi_{s,t}(x)$, $s, t \in R_+$, $x \in R^n$ be a continuous R^n-valued random field defined on the probability space $(\Omega, \mathcal{F}, P)$. It is called a *stochastic flow of homomorphisms* if it satisfies the following properties:

(1) $\varphi_{s,u} = \varphi_{t,u} \circ \varphi_{s,t}$ for any s, t, u almost surely where $\circ$ denotes the composition of maps;

(2) $\varphi_{s,s}$ = identity map for any s almost surely;

(3) $\varphi_{s,t}: R^n \to R^n$ is an onto homomorphism for any s, t almost surely.

It is called a *Brownian flow* if it still satisfies that

(4) For any $0 \leqq t_0 < t_1 < \cdots < t_k$, $\varphi_{t_i,t_{i+1}}$, $i = 0, \cdots, k-1$ are independent.

If $\varphi_{s,t}$ is only defined on $0 \leqq s \leqq t < \infty$, we call it a *forward stochastic flow of homomorphisms* or *forward Brownian flow*, respectively.

We shall assume the following conditions.

(A1) $\varphi_{s,t}(x)$ is square integrable for each s, t, x, and there exist the infinitesimal mean $b(x, t)$ and the infinitesimal covariance $a(x, y, t)$ for any t, x, y:

$$b(x, t) := \lim_{h \to 0+} \frac{1}{h} E\{\varphi_{t,t+h}(x) - x\},$$

$$a(x, y, t) := \lim_{h \to 0+} \frac{1}{h} E\{(\varphi_{t,t+h}(x) - x)(\varphi_{t,t+h}(y) - y)^T\}.$$

(A2) There exists a positive constant K such that

$$|E\{\varphi_{s,t}(x) - x\}| \leqq K(1 + |x|)|t - s|,$$

$$|E\{(\varphi_{s,t}(x) - x)(\varphi_{s,t}(y) - y)^T\}| \leqq K(1 + |x|)(1 + |y|)|t - s|$$

for any s, t, x, y.

(A3) $a(x, y, t)$ and $b(x, t)$ are continuous in (x, y, t) and (x, t), respectively. Moreover, they are locally δ-Hölder continuous ($\delta > 0$): for any compact subset C of R^n, there exists a positive constant K_c such that

$$\|a(x, x, t) - 2a(x, y, t) + a(y, y, t)\| \leqq K_c|x - y|^{2\delta},$$

$$|b(x, t) - b(y, t)| \leqq K_c|x - y|^{\delta}$$

hold for any x, $y \in C$, and $t \geqq 0$.

We shall need the following theorem due to Kunita [9, Thm. 4.2.8].

THEOREM 4.1 (Kunita [9]). *Let $\varphi_{s,t}(x)$, $0 \leqq s \leqq t < \infty$, $x \in R^n$ be a forward Brownian flow. Suppose that the pair of infinitesimal covariance and infinitesimal mean $(a(x, y, t), b(x, t))$ satisfies* (A1)-(A3). *Then there exists a Brownian motion $F(x, t)$ such that the flow is governed by Itô's stochastic differential equation based on $F(x, t)$, i.e.,*

$$\varphi_{s,t}(x) = x + \int_s^t F(\varphi_{s,r}(x), dr) \quad a.s. \quad on\ 0 \leqq s \leqq t < \infty, \quad x \in R^n.$$

Furthermore, the mean and the covariance of $F(x, t)$ coincide with

$$\int_0^t b(x, r)\, dr \quad and \quad \int_0^t a(x, y, r)\, dr,$$

respectively.

We now have the following theorem immediately which shows we can use the Lyapunov function to estimate the Lyapunov exponent of the stochastic flow.

THEOREM 4.2. *Let $\varphi_{s,t}(x)$, $0 \leqq s \leqq t < \infty$, $x \in R^n$ be a forward Brownian flow. Suppose that the pair of infinitesimal covariance and infinitesimal mean $(a(x, y, t), b(x, t))$ satisfies* (A1)-(A3). *Assume furthermore that there exist a function $V \in C^2(R^n)$, a polynomial $\mu(t)(t \geqq 0)$ with positive coefficients, and constants $p > 0$, $\lambda > 0$, and $\rho \geqq 0$ such that*

(1) $|x|^p \leqq V(x)$, $x \in R^n$;

(2) $LV(x) \leqq -\lambda V(x) + \mu(t)\, e^{-\lambda t + \rho t}$, $(x, t) \in R^n \times R_+$;

$$(3) \qquad \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} V(x) \frac{\partial}{\partial x_j} V(x) a^{ij}(x, x, t) \leqq \mu(t)\, e^{-\lambda t + \rho t} V(x), \qquad (x, t) \in R^n \times R_+.$$

Then the flow satisfies

$$\limsup_{t \to \infty} \frac{1}{t} \log |\varphi_{s,t}(x)| \leqq -(\lambda - \rho)/p \quad a.s.$$

for any $s \geqq 0$ and $x \in R^n$.

For a stochastic flow of homomorphisms we can have the similar result, and we leave the details to the readers.

5. Examples. In this section we shall give some simple examples to illustrate our results.

Example 5.1. Let $w(\cdot)$ be a one-dimensional Wiener process. Consider an Itô equation

$$(5.1) \qquad dx(t) = -\lambda(2 - \sin t)x(t)\, dt + p(t)\, e^{-\gamma t}\, dw(t) \quad \text{on } t \geqq 0,$$

where λ, γ are positive constants and $p(t)$ is a polynomial of t. Define a Lyapunov function $V(x) = x^2$. We have

$$LV(x) = -2\lambda(2 - \sin t)x^2 + p(t)^2\, e^{-2\gamma t} \leqq -2(\lambda \wedge \gamma)V(x) + p(t)^2\, e^{-2(\lambda \wedge \gamma)t}$$

and

$$V_x(x)^2 p(t)^2\, e^{-2\gamma t} \leqq 4p(t)^2\, e^{-2(\lambda \wedge \gamma)t} V(x).$$

Hence, by Corollary 3.3, the solution of (5.1) satisfies

$$\limsup_{t \to \infty} \frac{1}{t} \log |x(t)| \leqq -(\lambda \wedge \gamma) \quad \text{a.s.}$$

Example 5.2. We still let $w(\cdot)$ be a one-dimensional Wiener process. Consider a linear stochastic oscillator

$$(5.2) \qquad \ddot{x}(t) + 3\dot{x}(t) + 2x(t) = p(t)\, e^{-\rho t} \dot{w}(t) \quad \text{on } t \geqq 0,$$

where $p(t)$ is a polynomial of t and ρ is a positive constant. Set $y = \dot{x}$, and the corresponding Itô stochastic differential equation is

$$d\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} dt + \begin{bmatrix} 0 \\ p(t)\, e^{-\rho t} \end{bmatrix} dw(t).$$

Define a Lyapunov function $V(x, y) = 8x^2 + 4xy + 2y^2$. We then have

(1) $V(x, y) \geqq x^2 + y^2$,

(2) $LV(x, y) = -8x^2 - 4xy - 8y^2 + p(t)^2\, e^{-2\rho t} \leqq -V(x, y) + p(t)^2\, e^{-2\rho t}$,

(3) $V_y(x, y)^2 p(t)^2\, e^{-2\rho t} \leqq 4p(t)^2\, e^{-2\rho t} V(x, y)$.

Hence, by Corollary 3.3, the solution of (5.2) satisfies

$$\limsup_{t \to \infty} \frac{1}{t} \log [x^2(t) + y^2(t)] \leqq -(1 \wedge 2\rho) \quad \text{a.s.}$$

Example 5.3. We finally consider a two-dimensional stochastic differential equation

$$\varphi_t = x + \int_0^t F(\varphi_s, ds), \tag{5.3}$$

where $F(x, t)$ is a continuous C-semimartingale with spatial parameter x. Suppose $F(x, t)$ has the characteristic $(a(x, y, t), b(x, t), A_t)$ such that $t \leqq A_t \leqq 2t$ for all $t \geqq 0$ and

$$b(x, t) = \begin{bmatrix} x_1 - 2x_2 \\ 3x_1 - 4x_2 \end{bmatrix}, \qquad a(x, x, t) = t^2 e^{-2t} \begin{bmatrix} \sin^2 x_1, & \sin x_1 \sin (x_1 + x_2) \\ \sin x_1 \sin (x_1 - x_2), & \sin^2 (x_1 - x_2) \end{bmatrix}.$$

Using a Lyapunov function

$$V(x) = \tfrac{1}{3}(13x_1^2 - 14x_1x_2 + 4x_2^2)$$

we have

(1) $V(x) \geqq |x|^2$,

(2) $LV(x) \leqq \frac{1}{3}(-16x_1^2 + 14x_1x_2 - 4x_2^2) + p(t)\, e^{-t} \leqq -V(x) + p(t)\, e^{-2t}$,

(3) $$\sum_{i,\, j=1}^{2} \frac{\partial}{\partial x_i} V(x) \frac{\partial}{\partial x_j} V(x) a^{ij}(x, x, t) \leqq p(t)\, e^{-2t},$$

where $p(t)$ is a polynomial of t. Thus, by Theorem 2.1, the solution of (5.3) satisfies

$$\limsup_{t \to \infty} \frac{1}{t} \log |\varphi_t| \leqq -\frac{1}{2} \quad \text{a.s.}$$

Acknowledgments. I thank Professors K. D. Elworthy, L. Markus, and J. Zabczyk for their helpful discussions. Thanks are also due to the referee for his useful suggestions.

REFERENCES

[1] L. ARNOLD, *A formula connectir sample and moment stability of linear stochastic systems*, SIAM J. Appl. Math., 44 (1984), pp. 793-802.

[2] L. ARNOLD AND W. KLIEMANN, *Qualitative theory of stochastic systems*, in Probabilistic Analysis and Related Topics, Vol. 3, A. T. Bharucha-Reid, ed., New York, Academic Press, 1983, pp. 1-79.

[3] L. ARNOLD, E. OELJEKLAUS, AND E. PARDOUX, *Almost sure and moment stability for linear Ito equations*, Lecture Notes in Mathematics, Vol. 1186, Springer-Verlag, Berlin, New York, 1984, pp. 129-159.

[4] A. CARVERHILL, *Flows of stochastic dynamical systems: ergodic theory*, Stochastics, 14 (1985), pp. 273-317.

[5] M. CHAPPELL, *Bounds for average Lyapunov exponents of gradient stochastic systems*, Lecture Notes in Mathematics, Vol. 1186, Springer-Verlag, Berlin, New York, 1984, pp. 292-307.

[6] H. CRAUEL, *Lyapunov numbers of Markov solutions of linear stochastic systems*, Stochastics, 14 (1984), pp. 11-28.

[7] R. CURTAIN, ED., *Stability of stochastic dynamical systems*, Lecture Notes in Mathematics, Vol. 294, Springer-Verlag, Berlin, New York, 1972.

[8] R. Z. HAS'MINSKII, *Necessary and sufficient conditions for the asymptotic stability of linear stochastic systems*, Theory Probab. Appl., 12 (1967), pp. 144-147.

[9] H. KUNITA, *Stochastic flows and stochastic differential equations*, to appear.

[10] X. MAO, *Lebesgue-Stieltjes integral inequality and stochastic stability*, Quart. J. Math. Oxford, 40 (1989), pp. 301-311.

[11] X. MAO AND R. WU, *Existence and uniqueness of the solutions of stochastic differential equations*, Stochastics, 11 (1983), pp. 19-32.

[12] G. S. LADDE, *Stochastic stability analysis of model ecosystems with time-delay*, in Differential Equations and Applications in Biology, Epidemics, and Population Problems, S. N. Busenberg and K. Cook, eds., Academic Press, New York, 1981, pp. 215-228.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 6, pp. 1491-1503, November 1990

013

GLOBAL STABILIZATION OF PARTIALLY LINEAR COMPOSITE SYSTEMS*

A. SABERI†, P. V. KOKOTOVIC‡, AND H. J. SUSSMANN§

Abstract. A linear stabilizable, nonlinear asymptotically stable, cascade system is globally stabilizable by smooth dynamic state feedback if (a) the linear subsystem is right invertible and weakly minimum phase, and, (b) the only linear variables entering the nonlinear subsystem are the output and the zero dynamics corresponding to this output. Both of these conditions are coordinate-free and there is freedom of choice for the linear output variable. This result generalizes several earlier sufficient conditions for stabilizability. Moreover, the weak minimum-phase condition for the linear subsystem cannot be relaxed unless a growth restriction is imposed on the nonlinear subsystem.

Key words. composite systems, stabilization, Lyapunov function, nonlinear control

AMS(MOS) subject classifications. 93C10, 93C15, 93A20

1. Introduction. In this paper we propose new sufficient conditions for global stabilization, by means of state feedback, of *composite partially linear systems* in the form

$$\dot{x} = f(x, \xi), \qquad x \in R^n, \quad \xi \in R^q, \tag{1.1a}$$

$$\dot{\xi} = A\xi + Bu, \qquad u \in R^m, \tag{1.1b}$$

where $f(x, \xi)$ is a smooth (i.e., C^∞) function and A and B are constant matrices. Throughout the paper it is assumed that:

(H1) The pair (A, B) is stabilizable.

(H2) The equilibrium $x = 0$ of $\dot{x} = f(x, 0)$ is globally asymptotically stable (GAS) and a smooth Lyapunov function $V(x) > 0$, $x \neq 0$; $V(0) = 0$, is known such that $V(x) \to \infty$ as $\|x\| \to \infty$

and

$$\nabla V(x) f(x, 0) < 0 \quad \text{for all } x \neq 0. \tag{1.2}$$

As a class of nonlinear composite systems [13], [19], [26], the partially linear systems (1.1) have become prominent because of recent results on partial feedback linearization, where $\dot{x} = f(x, 0)$ is referred to as the "nonlinear zero dynamics" [2]-[4], [9], [12]. It would appear that when $x = 0$ is globally asymptotically stable as assumed by (H2), then the global stabilization of the whole system should not be difficult. Simple examples show that this is not so. Disturbed by an exponentially decaying input $\xi(t)$, the nonlinear system (1.1a) can become unstable, or even worse: its state may escape to infinity in finite time! One way to circumvent this difficulty is to restrict $f(x, \xi)$ by a global linear growth condition and then to apply the classical "total stability" theorems [7]. A criticism of the global linear growth assumption is that it does not let nonlinear systems be "nonlinear enough." It excludes simple chemical kinetics, mechanical phenomena such as Coriolis forces, etc.

* Received by the editors July 24, 1989; accepted for publication December 4, 1989.

† Department of Electrical and Computer Engineering, Washington State University, Pullman, Washington 99164-2752. The work of this author was supported by National Science Foundation grant ECS-8618953 and by Boeing Commercial Airplane Group.

‡ Coordinated Science Laboratory, University of Illinois, 1101 West Springfield Avenue, Urbana, Illinois 61801. The work of this author was supported by United States Department of Energy grant DE-FGD2-88ER13939 and National Science Foundation grant ECS-8818166.

§ Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903. The work of this author was supported in part by National Science Foundation grant DMS-83-01678-01.

This paper continues the efforts of several recent studies [5], [10], [25], [22], [11], which do not make a linear growth assumption. Instead of constraining the nonlinear nature of $f(x, \xi)$, our approach characterizes its dependence on ξ by the expression

$$f(x, \xi) - f(x, 0) = G(x, \xi)C\xi. \tag{1.3}$$

Since for a given $f(x, \xi)$ the choice of G and C is not unique, we seek a smooth $n \times p$ matrix function $G(x, \xi)$ and a constant $p \times q$ matrix C to encompass the largest class of linear systems

$$\dot{\xi} = A\xi + Bu, \qquad \xi \in R^q, \quad u \in R^m, \tag{1.4a}$$

$$y = C\xi, \qquad y \in R^p. \tag{1.4b}$$

Our main result is a stable right invertibility (SRI) condition imposed on (1.4). This condition encompasses a much broader class of systems than the feedback positive real (FPR) condition of [11]. When the linear subsystem is not SRI, our second result imposes a restriction on the nonlinear subsystem, which is less severe than the linear growth condition.

The meaning of (H1) and (H2) is that each subsystem, taken isolated, is globally stable (or stabilizable). This setting is suitable for the construction of composite Lyapunov functions [13], [19], which we use to broaden the class of linear subsystems (1.4). In § 2 we start with a sum-composite Lyapunov function, leading to the class of stable invertible systems of relative degree one (SI_1). This class includes the FPR systems of [11], and is broadened by the assumption that the zero dynamics are stable ("weak minimum phase"), rather than asymptotically stable ("minimum phase"). The main result of § 3, and of the whole paper, removes the relative degree assumption and requires only that the linear subsystem (1.4) be stable right invertible. The analysis leading to this result provides new insights into linear system properties, revealed by the special coordinate basis (s.c.b.) of [17] and [23], which is our key analytical tool. As shown in § 4, the assumptions of the main theorem cannot be weakened unless some additional restrictions are imposed on $f(x, \xi)$. So, when the linear subsystem (1.4) is not SRI and the results of §§ 2 and 3 are not applicable, then § 4 introduces a constraint on the nonlinear subsystem.

2. The stabilization procedure in the case SI_1. The problem is to find a smooth feedback control

$$u = K\xi + v(x, \xi), \tag{2.1}$$

which guarantees the GAS property for the equilibrium $(x, \xi) = (0, 0)$ of the feedback system

$$\dot{x} = f(x, 0) + G(x, \xi)C\xi, \tag{2.2a}$$

$$\dot{\xi} = (A + BK)\xi + Bv(x, \xi). \tag{2.2b}$$

This system is obtained by applying the control (2.1) to the system (1.1) and taking into account the representation (1.2) of $f(x, \xi)$. The two subsystems clearly displayed in (2.2) are

$$\dot{x} = f(x, 0), \tag{2.3a}$$

$$\dot{\xi} = (A + BK)\xi \triangleq A_K\xi. \tag{2.3b}$$

By (H2) a Lyapunov function for the nonlinear subsystem (2.3a) is $V(x)$, while (H1) assures the existence of K such that $\operatorname{Re}\lambda(A_k)<0$. Hence, a Lyapunov function for the linear subsystem (2.3b) can be chosen as $\xi^T P\xi$, where $P=P^T>0$ is such that

$$PA_K + A_K^T P = -Q \leqq 0, \tag{2.4a}$$

$$(Q^{1/2}, A_K) \text{ detectable.} \tag{2.4b}$$

Our approach is to use $V(x)$ and $\xi^T P\xi$ to form a composite Lyapunov function $W(x, \xi)$ for the whole system (2.2). The simplest choice is

$$W(x, \xi) = V(x) + \xi^T P\xi. \tag{2.5}$$

Its derivative for (2.2) is

$$\dot{W}(x, \xi) = \nabla V(x)[f(x, 0) + G(x, \xi)C\xi] - [\xi^T Q\xi - 2\xi^T PBv(x, \xi)]. \tag{2.6}$$

This expression is not informative because it contains the interconnection terms which are sign indefinite. However, if $G(x, \xi)$, C, P, and $v(x, \xi)$ can be found such that the interconnection terms in (2.6) are cancelled, then

$$\dot{W}(x, \xi) = \nabla V(x) f(x, 0) - \xi^T Q\xi. \tag{2.7}$$

A sufficient condition for being able to achieve the cancellation is

$$B^T P = C. \tag{2.8}$$

Under this condition, the explicit form of v resulting in (2.7) is

$$v(x, \xi) = -\tfrac{1}{2}[\nabla V(x) G(x, \xi)]^T. \tag{2.9}$$

Remark 1. Assuming, without loss of generality, that B and C are of full rank, (2.8) implies the same number of inputs and outputs $p=m$. This restriction will be removed in § 3.

PROPOSITION 1. *Suppose there exists a K such that* (2.4) *and* (2.8) *are satisfied. Then the equilibrium* $(x, \xi)=(0, 0)$ *of the system* (2.2) *with this K and* (2.9) *is* GAS.

Proof. It is clear from (2.7) that $\dot{W}(x, \xi) \leqq 0$ for all (x, ξ) and $\dot{W}(x, \xi)<0$ if $x \neq 0$. Moreover, $W(x, \xi) \geqq 0$ for all x and ξ and equality holds if and only if $(x, \xi)=(0, 0)$. This establishes global stability of $(x, \xi)=(0, 0)$, since $W(x, \xi) \to \infty$ as $\|(x, \xi)\| \to \infty$. To establish the GAS of the $(x, \xi)=(0, 0)$ it suffices to show that, if $\gamma: t \to (x(t), \xi(t))$ is a complete trajectory of (2.2) along which $\dot{W}=0$, then it follows that $x(t) \equiv 0$ and $\xi(t) \equiv 0$. To begin with, $x(t)$ must be zero for all t, because $\dot{W}(x, \xi)<0$ unless $x=0$. Moreover, $x(t) \equiv 0$ implies that v defined by (2.9) vanishes along γ. Therefore, $t \to \xi(t)$ is a solution of $\dot{\xi}=A_K\xi$ and $\dot{W}(x, \xi) = -\xi^T(t)Q\xi(t) = 0$ for all t. By the detectability assumption (2.4b) this implies $\xi(t) \equiv 0$ and, hence, $(x, \xi)=(0, 0)$ is GAS. □

The above construction is a variant of the cancellation procedure used in the model reference adaptive control and goes back to [16] and [15].

With Proposition 1 the stabilization problem is reduced to that of the existence of a K satisfying (2.4) and (2.8). In [11] this issue was addressed indirectly, via a positive real property of (C, A_K, B). Here we will deal directly with the properties of the linear subsystem (1.4) induced by (2.4) and (2.8), such as invertibility, relative degree, and zero dynamics. Let us recall their definitions.

Invertibility. The linear system (1.4) is said to be invertible if, for any C^q function $y_{\text{ref}}(t)$, where q is an integer, there exist $u(t)$ and $\xi(0)$ such that $y(t)=y_{\text{ref}}(t)$ for all $t \in [0, \infty)$.

Relative degree. When (1.4) is "square," $p = m$, it is said to have scalar relative degree r if its first $r-1$ Markov parameters are zero, $CA^iB = 0$ for $i = 0, 1, \cdots, r-2$, and $CA^{r-1}B$ is nonsingular. Equivalently, the system (1.4) has relative degree r if it is invertible and all of its infinite zeros are of order r.

Zero dynamics. Let V^* be the supremal (A, B)-invariant subspace in Ker C, and let R^* be the supremal (A, B)-controllability subspace in Ker C. The solutions $\xi(t)$ of (1.4) restricted for all $t \in [0, \infty]$ to V^*/R^* are called the zero dynamics of (1.4). When (1.4) is invertible its zero dynamics are equivalently defined as the solutions $\xi(t)$ satisfying $y(t) \equiv 0$ for all t.

Weak minimum phase. An invertible linear system (1.4) is said to be weak minimum phase, or, equivalently, stable invertible (SI), if its zero dynamics are stable in the sense of Lyapunov.

We are now in the position to completely characterize the class of linear systems (1.4) specified by (2.4) and (2.8).

PROPOSITION 2. *The following two statements are equivalent*:

(a) *For the system* (1.4) *there exists K satisfying* (2.4) *and* (2.8).

(b) *The system* (1.4) *is stabilizable, stable invertible and, moreover, its leading Markov parameter CB is symmetric positive definite.*

Proof. (a) $\rightarrow$ (b). We postmultiply (2.8) by B and verify that $CB = B^TC^T > 0$. Hence, (1.4) is invertible and has relative degree one. To prove the stable invertiblity (weak minimum phase) property of (1.4) we assume, without loss of generality, that (1.4) is in the special coordinate basis (s.c.b.)

$$\dot{\xi}_0 = A_0\xi_0 + A_1\xi_1, \tag{2.10a}$$

$$\dot{\xi}_1 = D_0\xi_0 + D_1\xi_1 + CBu, \tag{2.10b}$$

$$y = \xi_1. \tag{2.10c}$$

This s.c.b. has evolved from early works [20], [14], and [6] and its general form is given in [17] and [23]. Noting that CB is nonsingular, the choice of u to achieve $\dot{\xi}_1 = 0$ for all t is obvious from (2.10b). With this choice, $\xi_1(0) = 0$ implies $y(t) = \xi_1(t) \equiv 0$ for all $t \in [0, \infty)$, so that the zero dynamics of (2.10) are the solutions of

$$\dot{\xi}_0 = A_0\xi_0. \tag{2.11}$$

Hence, the eigenvalues of A_0 are the invariant zeros of (2.10). A simple calculation reveals an important property induced by the cancellation condition (2.8). Under this condition, P for the system (2.10) is block diagonal, $P = \text{diag}\,(P_0, P_1)$, where P_0 and P_1 are positive-definite matrices of dimensions $(q-m)\times(q-m)$ and $m \times m$, respectively. Because of this property and using any $K = (K_0, K_1)$ appropriately partitioned, the Q matrix in (2.4) is of the form

$$Q = -\begin{pmatrix} P_0A_0^T + A_0P_0 & * \\ * & * \end{pmatrix}. \tag{2.12}$$

By assumption (a) this matrix is positive semidefinite, which implies (see [1]) that

$$P_0A_0 + A_0^TP_0 \leqq 0. \tag{2.13}$$

Thus the zero dynamics are stable, which completes the proof of (a) $\rightarrow$ (b).

(b) $\rightarrow$ (a) Since the system is invertible and has relative degree one, we can represent it by (2.10). Moreover, the stable-invertibility assumption implies that the zero dynamics system (2.11) is stable. Without loss of generality we now let $A_0 = \text{diag}\,(A_{01}, A_{02})$, where

$$\text{Re}\,\lambda(A_{01}) < 0, \quad \text{Re}\,\lambda(A_{02}) = 0, \quad \text{and} \quad A_{02} + A_{02}^T = 0. \tag{2.14}$$

Then the system (2.10) is rewritten as

$$\dot{\xi}_{01} = A_{01}\xi_{01} + A_{11}\xi_1, \tag{2.15a}$$

$$\dot{\xi}_{02} = A_{02}\xi_{02} + A_{12}\xi_1, \tag{2.15b}$$

$$\dot{\xi}_1 = D_{01}\xi_{01} + D_{02}\xi_{02} + D_1\xi_1 + CBu, \tag{2.15c}$$

$$y = \xi_1. \tag{2.15d}$$

The Hurwitz property of A_{01} allows us to define $P_{01} = P_{01}^T > 0$ as the solution of

$$P_{01}A_{01} + A_{01}^T P_{01} = -I. \tag{2.16}$$

To prove the existence of K satisfying (2.4) and (2.8) we make a particular choice of $K = (K_{01}, K_{02}, K_1)$:

$$K = -[(CB)^{-1}D_{01} + A_{11}^T P_{01}, (CB)^{-1}D_{02} + A_{12}^T, (CB)^{-1}D_1 + \tfrac{1}{2}I]. \tag{2.17}$$

For this choice of K, the matrix A_K for (2.15) is

$$A_K = \begin{pmatrix} A_{01} & 0 & A_{11} \\ 0 & A_{02} & A_{12} \\ -(CB)A_{11}^T P_{01} & -(CB)A_{12}^T & -\tfrac{1}{2}CB \end{pmatrix}. \tag{2.18}$$

The substitution of this A_K and $P = \operatorname{diag}[P_{01}, I, (CB)^{-1}]$ into (2.4a) and (2.8) proves that they satisfy (2.4a) and (2.8) with $Q = \operatorname{diag}(I, 0, I)$. To prove that (2.4b) is also satisfied we use $Q^{1/2} = \left(\begin{smallmatrix} I & 0 & 0 \\ 0 & 0 & I \end{smallmatrix}\right)$ and test the observability of the pair $(Q^{1/2}, A_K)$. The stabilizability of (A, B), assumption (H1), implies the controllability of (A_{02}, A_{12}), and, hence, the matrix $[sI - A_{02}, A_{12}]$ is full rank for all complex s. It follows that

$$\operatorname{rank}\begin{bmatrix} Q^{1/2} \\ sI - A_K \end{bmatrix} = q \quad \text{for all complex } s. \tag{2.19}$$

Thus $(Q^{1/2}, A_K)$ is observable and (2.4b) is satisfied. □

Applying Propositions 1 and 2 to our stabilization problem we summarize the results of this section as follows.

THEOREM 1. *Suppose that for the composite system* (1.1), *with f represented by* (1.2), (H1) *and* (H2) *hold, and the linear subsystem* (1.4) *is invertible with relative degree one and weakly minimum phase* (SI_1). *Then there exists a feedback law such that the equilibrium* $(x, \xi) = (0, 0)$ *of the closed-loop system* (2.2) *is* GAS. *A particular form of this feedback law is* (2.1) *with* $v(x, \xi)$ *given by* (2.9) *and* K *given by* (2.17).

Proof. In the (SI_1) systems the matrix CB is nonsingular, while in Proposition 2 it is assumed that CB is symmetric positive definite. However, it follows from (2.10b) that, with a static precompensator $\bar{u} = (CB)^{-1}u$, both Propositions 1 and 2 are applicable to any (SI_1) system. Alternatively, the same effect can be achieved with the postcompensator $\bar{y} = (CB)^{-1}y$. □

A question raised by the following example is whether the weak minimum-phase condition required in Theorem 1 is in some sense necessary.

Example 1. For $c_2 > 0$ the linear subsystem in

$$\dot{x} = -x^3 - x^3 y, \quad \dot{\xi}_1 = \xi_2, \quad \dot{\xi}_2 = u, \tag{2.20a}$$

$$y = c_1\xi_1 + c_2\xi_2 \tag{2.20b}$$

is invertible with relative degree one, and (H1), (H2) are satisfied. For $c_1 \geqq 0$ the weak minimum-phase assumption is satisfied and, by Theorem 1, the system (2.20) is globally

asymptotically stabilizable. What if the weak minimum-phase condition of Theorem 1 is not satisfied, that is, $c_1 < 0$? Then, as shown in [11], for all initial states $(x_0, \xi_{10}, \xi_{20})$ such that

$$(1 - c_1\xi_{10})x_0^2 > \frac{c_1}{2c_2} \tag{2.21}$$

the system (2.19) fails to be asymptotically controllable to zero and therefore fails to be smoothly stabilizable.

We will return to this issue in § 4 and show that the weak minimum-phase condition is necessary in the sense that, in general, it cannot be weakened without a further restriction on $f(x, \xi)$.

3. The stabilization procedure in the case SRI. The first generalization of Theorem 1 and, at the same time, a step toward our main result, is a global stabilization condition for the system

$$\dot{x} = f(x, 0) + G(x, \xi_0, \xi_1)\xi_1, \qquad x \in R^n, \tag{3.1a}$$

$$\dot{\xi}_0 = A_0\xi_0 + A_1\xi_1, \qquad \xi_0 \in R^{q_0}, \tag{3.1b}$$

$$\dot{\xi}_1 = \xi_2, \qquad \xi_i \in R^m, \quad i = 1, \cdots, r, \tag{3.1c}$$

$$\vdots$$

$$\dot{\xi}_{r-1} = \xi_r, \qquad \xi \in R^q, \quad q = q_0 + rm, \tag{3.1d}$$

$$\dot{\xi}_r = u_r, \quad y = \xi_1, \quad u_r \in R^n, \quad y \in R^m. \tag{3.1e}$$

The linear part of this system is in the form to which every invertible relative degree r (SI_r) system (1.4) can be transformed using first the s.c.b. of [17], [23], and [24] and then a feedback transformation $u = (CA^{r-1}B)^{-1}(Fx + u_r)$, with an appropriate F. The zero dynamics of this linear system are defined by (3.1b) with $\xi_1 = 0$, and the weak minimum phase property (SI_r) implies that they are stable. To simplify notation, we assume that A_0 does not have an asymptotically stable part, i.e., we let

$$A_0^T + A_0 = 0. \tag{3.2}$$

There is no loss of generality here because if some of the linear zero dynamics are asymptotically stable, we simply incorporate them in the nonlinear subsystem (3.1a) with an obvious redefinition of x, f, and G. However, our next assumption, already satisfied by the special form of (3.1a), is essential.

(H3) In (1.1) the dependence of $f(x, \xi)$ on ξ is such that (1.3) has the form

$$f(x, \xi) - f(x, 0) = G(x, \xi_0, \xi_1)\xi_1, \tag{3.3}$$

that is, G is allowed to depend only on the output $y = \xi_1$ and the linear zero dynamics ξ_0 induced by this output.

This assumption is a structural characterization of the linear/nonlinear interconnection (1.3). A choice of $y = Cx = \xi_1$ uniquely specifies ξ_0 via its s.c.b. Then (3.3) may or may not be satisfied even when (1.3) is satisfied. Let us illustrate this point.

Example 2. For the system

$$\dot{x} = -x^3 - \xi_1(\alpha\xi_1 + \xi_2)x^3, \qquad \dot{\xi}_1 = \xi_2, \quad \dot{\xi}_2 = u \tag{3.4}$$

the choice of G and C in (3.3) depends on α. If $\alpha \geqq 0$, then the choice $y = \alpha\xi_1 + \xi_2$ results in a linear stable invertible system with $r = 1$ so that Theorem 1 applies. If $\alpha < 0$

then the same linear system is nonminimum phase and Theorem 1 does not apply. So we must try the second choice $y = \xi_1$, resulting in a linear system with $r = 2$ and trivially minimum phase, because it has no finite zeros. However, now the connection structure condition (H3) is not satisfied because $G = (\alpha\xi_1 + \xi_2)x^3$ depends not only on ξ_1, but also on ξ_2. In Example 3 we will discuss an important implication of this violation of (H3).

Returning to (3.1) let us recall from Theorem 1 and (2.17) that for the case $r = 1$ a stabilizing control for (3.1) with (3.2) is

$$u_1(x, \xi_0, \xi_1) = -A_1^T \xi_0 - \tfrac{1}{2}\xi_1 + v(x, \xi). \tag{3.5}$$

With these preliminaries out of the way, the stabilization condition or the case of relative degree r is obtained using the chain of integrators result [11], [10], [25], [22].

PROPOSITION 3. *Suppose that the composite system* (1) *satisfies* (H1) *and* (H2) *and that the linear subsystem* (1.4) *is invertible with relative degree r and weakly minimum phase* (SI_r). *If, in addition, the connection-structure condition* (H3) *is satisfied, then this composite system is globally asymptotically stabilizable at* $(x, \xi) = (0, 0)$ *by a smooth state feedback control. Furthermore, the expressions for a stabilizing control and for a corresponding Lyapunov function can be derived recursively.*

Proof. It is sufficient to prove this statement for the system (3.1). Let us start with the case $r = 2$. From the first three equations (3.1a)–(3.1c) the result would be known from Theorem 1, if ξ_2 were the control variable u_1 in (3.5). This suggests that ξ_2 be modified as follows:

$$\xi_2 = u_1(x, \xi_0, \xi_1) + \tilde{\xi}_2, \qquad \tilde{\xi}^T = [\xi_0^T, \xi_1^T, \tilde{\xi}_2^T]. \tag{3.6}$$

The time derivative of u_1 along the solutions of (3.1) can be evaluated explicitly as a function of x and $\tilde{\xi}$. We denote it by

$$\left.\frac{du_1}{dt}\right|_{(3.1)} = h_1(x, \tilde{\xi}). \tag{3.7}$$

Then for $r = 2$ the system (3.1) becomes

$$\dot{x} = f(x, 0) + G(x, \xi_0, \xi_1)\xi_1, \tag{3.8a}$$

$$\dot{\xi}_0 = A_0\xi_0 + A_1\xi_1, \tag{3.8b}$$

$$\dot{\xi}_1 = \tilde{\xi}_2 + u_1(x, \xi_0, \xi_1), \tag{3.8c}$$

$$\dot{\tilde{\xi}}_2 = -h_1(x, \xi_0, \xi_1) + u_2. \tag{3.8d}$$

For this system we use the Lyapunov function

$$W_2(x, \tilde{\xi}) = V(x) + \|\tilde{\xi}\|^2. \tag{3.9}$$

Its time derivative for (3.8) is

$$\dot{W} = \nabla V(x) f(x, 0) - \|\xi_1\|^2 + 2\tilde{\xi}_2^T[\xi_1 - h_1(x, \tilde{\xi}) + u_2]. \tag{3.10}$$

An obvious choice of u_2 that makes $\dot{W} \leqq 0$ is

$$u_2(x, \tilde{\xi}) = -\xi_1 - \tfrac{1}{2}\tilde{\xi}_2 + h_1(x, \tilde{\xi}). \tag{3.11}$$

The remaining step of the proof that $(x, \xi_0, \xi_1, \tilde{\xi}_2) = (0, 0, 0, 0)$ is the GAS equilibrium of (3.8) is, as in Proposition 2, via an observability property which is guaranteed by the controllability of (A_0, A_1). The return to the original coordinates via (3.6) shows that $\tilde{\xi}_2 \to 0$ implies $\xi_2 \to 0$, which completes the proof for $r = 2$.

To proceed to the case $r=3$ we note that, if ξ_3 were the control variable, the result (3.11) for $r=2$ would apply, which in turn suggests the modification

$$\xi_3 = u_2(x, \xi_0, \xi_1, \xi_2) + \tilde{\xi}_3 \tag{3.12}$$

where u_2 is expressed using ξ_2 rather than $\tilde{\xi}_2$. Adding the term $\|\tilde{\xi}_3\|^2$ to W_2 the new Lyapunov function W_3 is formed. Requiring that $\dot{W}_3 \leqq 0$ we obtain a stabilizing $u_3(x, \tilde{\xi})$ for the case $r=3$. It is clear that this procedure can be continued for any r, which completes the proof. □

Once again, an example is used to illustrate the closeness of the sufficient condition above to being also necessary.

Example 3. Let us reexamine the system (3.4) in Example 2 in the case when $\alpha<0$ and $y=\xi_1$. In this case $r=2$, but the connection structure (H3) is violated and Proposition 3 does not apply. A detailed calculation in [11] shows that in this case there are initial conditions $\{x(0), \xi_1(0), \xi_2(0)\}$ for which the solutions of (3.4) are either unbounded as $t\to\infty$ or escape to infinity in finite time. It follows that for the system (3.4) the assumption (H3) cannot be relaxed to allow G to depend on both ξ_1 and ξ_2.

We are now prepared to remove the assumption that the linear system is "square," that is, $m=p$, and with a scalar relative degree. In the next step we allow $m \geqq p$ and require that the linear subsystem be right invertible and weakly minimum phase. The definitions of right invertibility and weak minimum phase are the same as in § 2 except that now we have $m \geqq p$. The problem of converting a right invertible system into an invertible one with scalar relative degree, which has been examined during the last two decades (e.g., [27], [21]), involves dynamic decoupling via precompensator and static feedback. In the following proposition, this conversion is achieved with the preservation of the weak minimum phase property using the results of [23] and [24].

PROPOSITION 4. *Consider the system* (1.4) *with* $m \geqq p$. *Assume that* (1.4) *is right invertible and let* $H(s)$ *be its transfer function matrix. Then there exists a precompensator* $u = C(s)\bar{u}$, $\bar{u} \in R^p$, *such that the system* $\bar{H}(s) \triangleq H(s)C(s)$ *has the following properties*:

(i) $\bar{H}(s)$ *has relative degree r.*

(ii) *Invariant zeros of* $\bar{H}(s)$ = *invariant zeros of* $H(s) \cup \Lambda$,

where Λ *denotes the set of additional invariant zeros induced by the compensator* $C(s)$ *and arbitrarily assignable.*

Proof. In the proof we construct two precompensators. The task of the first precompensator $u = C_1(s)\hat{u}$ is to "square down" $\hat{H}(s) \triangleq H(s)C_1(s)$ subject to the requirement that the "squared" system satisfies (ii). The task of the second precompensator $C_2(s)$ is that the compensated system $\bar{H}(s)$ be of relative degree r, but without changing the finite-zero structure of $\hat{H}(s)$. In other words, we require that invariant zeros of $\bar{H}(s)$ equal invariant zeros of $\hat{H}(s)$.

As the design of $C_1(s)$ was developed in [24], the remaining task is to design $C_2(s)$. Since $\hat{H}(s)$ is invertible, it can be represented in the s.c.b. of [23] as follows:

$$\dot{\xi}_0 = A_0\xi_0 + A_1\tilde{y}, \tag{3.12a}$$

$$\dot{\xi}_i = A_i\xi_i + B_i\left(\tilde{u}_i + \sum_{j=0}^{r} D_{ij}\xi_j\right) + L_i\tilde{y}, \qquad i = 1, \cdots, r, \tag{3.12b}$$

$$\tilde{y} = C_i\xi_i, \quad \tilde{y}_i \in R^{q_i}, \quad \tilde{y}^T = (\tilde{y}_1^T, \cdots, \tilde{y}_r^T), \quad \tilde{y} = \Gamma_1 y, \tag{3.12c}$$

$$\tilde{u}^T = (\tilde{u}_1^T, \cdots, \tilde{u}_r^T), \qquad \hat{u} = \Gamma_2\tilde{u}, \tag{3.12d}$$

where Γ_1, $\Gamma_2 \in R^{m\times m}$ are nonsingular matrices and

$$A_i = \begin{pmatrix} 0 & I_{(i-1)q_i} \\ 0 & 0 \end{pmatrix}, \quad B_i = \begin{pmatrix} 0 \\ I_{q_i} \end{pmatrix}, \quad C_i = (I_{q_i}, 0). \tag{3.13}$$

This s.c.b. displays the zero structure of the system:

Invariant zeros of $\hat{H}(s)$ = eigenvalues of A_0,
Zero dynamics of $\hat{H}(s)$ = the solutions of $\dot{\xi}_0 = A_0\xi_0$,
i = order of an infinite zero, iq_i = number of infinite zeros of order i.

Now, to design $C_2(s)$ that makes $\hat{H}(s)$ of relative degree r we simply add an appropriate number of integrators to each input $\tilde{u}_i$. Hence we let

$$\bar{u}^T = (\bar{u}_1^T, \cdots, \bar{u}_r^T), \quad \tilde{u}_i = \frac{1}{s^{r-i}}\bar{u}_i, \quad \tilde{u} = \tilde{C}_2(s)\bar{u}, \quad \tilde{C}_2(s) \triangleq \operatorname{diag}\left(\frac{1}{s^{r-i}}\right) \tag{3.14}$$

and obtain that $\hat{H}(s)\Gamma_2\tilde{C}_2(s)$ has relative degree r and its invariant zeros are the invariant zeros of $\hat{H}(s)$. So the second compensator is $C_2(s) \triangleq \Gamma_2\tilde{C}_2(s)$. □

Applying Propositions 3 and 4 to our stabilization problem we formulate the main result of this paper as follows.

THEOREM 2. *If the assumptions* (H1)-(H3) *hold, and the linear subsystem* (1.4) *is right invertible and weakly minimum phase, then the composite system* (1.1) *is globally asymptotically stabilizable at* $(x, \xi) = (0, 0)$ *by dynamic state feedback.*

4. Restrictions on the nonlinear part. An assumption made throughout this paper is that the full state of the composite system (1.1) is available for feedback. Despite this assumption, our stabilizability conditions impose restrictions on the input-output structure of the linear subsystem. In addition to the connection structure and right invertibility assumptions, the key restriction is that the linear subsystem be *weakly minimum phase.* The analysis of Example 1 has given us a hint that this key restriction is in some sense necessary. Pursuing this hint we now prove that, given a strictly nonminimum phase linear subsystem (1.4), a nonlinear subsystem can be found such that the cascade (1.1) of these two subsystems, satisfying (H1)-(H3), is not globally stabilizable. Our Theorem 3 reveals that the underlying instability mechanism is an interplay of unstable zero dynamics with rapidly growing nonlinear terms, such as x^3. To limit this interplay, in Proposition 5 we introduce a specific growth condition which is less restrictive than a global Lipshitz condition.

THEOREM 3. *Consider the composite system satisfying assumption* (H1)-(H3):

$$\dot{x} = f(x, 0) + G(x, \xi_0, y)y, \qquad x \in R^n, \tag{4.1a}$$

$$\dot{\xi} = A\xi + Bu, \qquad \xi \in R^q, \quad u \in R^m, \tag{4.1b}$$

$$y \in C\xi, \qquad y \in R^p, \tag{4.1c}$$

and let the dynamics of (4.1b), (4.1c) *associated with its invariant zeros be represented by*

$$\dot{\xi}_0 = A_0\xi_0 + A_1 y, \qquad \xi_0 \in R^q. \tag{4.2}$$

When (4.1b), (4.1c) *is strictly nonminimum phase, i.e., some of the eigenvalues of* A_0 *have positive real parts, then there exist* $f(x, 0)$ *and* $G(x, \xi_0, y)$ *satisfying* (H2) *and* (H3) *such that the composite system* (4.1) *is not globally stabilizable.*

Proof. Without loss of generality we assume that all the eigenvalues of A_0 are with positive real parts $\operatorname{Re}\lambda(A_0) > 0$. (If only some of them are, then we let $A_0 = \operatorname{diag}(A_{01}, A_{02})$, with $\operatorname{Re}\lambda(A_{02}) > 0$ and modify the proof to apply to the subsystems with A_{02} instead of with A_0.) Using the positive-definite P_0 satisfying

$$P_0A_0 + A_0^TP_0 = 2I, \tag{4.3}$$

we evaluate the derivative of $V_0 = \xi_0^T P_0 \xi_0$ along the trajectories of (4.2):

$$\begin{aligned} \dot{V}_0 &= 2\|\xi_0\|^2 + 2\xi_0^T P_0 A_1 y + \|P_0 A_1 y\|^2 - \|P_0 A_1 y\|^2 \\ &\geqq \|\xi_0\|^2 - \|P_0 A_1 y\|^2 \\ &\geqq \beta_1 V_0 - \beta_2 \|A_1 y\|^2 \end{aligned} \tag{4.4}$$

where $\beta_1 = 1/\lambda_{\max}(P_0)$, $\beta_2 = \|P_0\|^2$. We are now in the position to pick a nonlinear subsystem to complete the proof. Consider the nonlinear subsystem defined by

$$f(x, 0) = -x^3, \qquad G(x, \xi_0, y) = x^3 \beta_2 y^T A_1^T A_1, \tag{4.5}$$

which satisfies (H2), (H3) and is nontrivial because, in view of (H1), the pair (A_0, A_1) is completely controllable and, hence, $A_1 \neq 0$. Integrating the nonlinear subsystem

$$\dot{x} = -x^3 + \beta_2 y \|A_1 y\|^2 x^3 \tag{4.6}$$

we obtain

$$2x^2(t) = \left[\frac{1}{2x^2(0)} + t - \int_0^t \beta_2 \|A_1 y(s)\|^2 \, ds \right]^{-1} \triangleq \frac{1}{\theta(t)}. \tag{4.7}$$

Clearly, $\theta(0) > 0$ and $\theta(t)$ must remain nonnegative for all $t > 0$ or else $x(t)$ would escape to infinity. Thus, using (4.4) a necesary condition for $x(t)$ to remain bounded is

$$\frac{1}{2x^2(0)} + t + \int_0^t (\dot{V}_0(s) - \beta_1 V_0(s)) \, ds \geqq 0 \tag{4.8}$$

and, hence,

$$V_0(t) \geqq \beta_1 \int_0^t \left(V_0(s) - \frac{1}{\beta_1} \right) ds + V_0(0) - \frac{1}{2x^2(0)}. \tag{4.9}$$

Finally, applying a version of Gronwall's lemma, (4.9) implies that

$$V_0(t) \geqq \frac{1}{\beta_1} + \left(V_0(0) - \frac{1}{2x^2(0)} - \frac{1}{\beta_1} \right) e^{\beta_1 t}. \tag{4.10}$$

Now, from $V_0(0) = \xi^T(0) P_0 \xi(0)$ we observe that, for any given $x(0)$, there exists $\xi(0)$ such that the factor multiplying $e^{\beta_1 t}$ is positive and $V_0(t) = \xi_0^T(t) P_0 \xi_0(t)$ grows exponentially. This completes the proof, because $\theta(t) \geqq 0$, a necessary condition for boundedness of $x(t)$, implies that $\xi_0(t)$ grows unbounded as $t \to \infty$. For $\xi_0(t)$ to remain bounded, $\theta(t)$ must become negative at some finite time at which $x(t)$ escapes to infinity. □

While Theorem 3 shows the limits to stabilizability of the composite system caused by the nonminimum phase property of its linear part, the above proof reveals the underlying instability mechanism. The effort to stabilize the unstable linear zero dynamics may destabilize the nonlinear subsystem through some rapidly growing nonlinear connection terms. It is clear, therefore, that the class of nonlinear subsystems which can be cascaded with linear nonminimum phase subsystem must be restricted by restricting the growth of the connection terms. It turns out that, under one such restriction, the feedback loop needs to be closed only around the linear subsystem. With $u = K\xi$ and $v(x, \xi) = 0$, the feedback system (2.2) becomes

$$\dot{x} = f(x, 0) + G(x, \xi)\xi = f(x, \xi), \tag{4.11a}$$

$$\dot{\xi} = (A + BK)\xi = A_K \xi, \tag{4.11b}$$

where the decomposition of $f(x, \xi)$ in (4.11a) is always possible due to the smoothness of $f(x, \xi)$. The assumption (H2) is now strengthened by requiring global exponential stability (GES), rather than only global asymptotic stability of $\dot{x} = f(x, 0)$. Another crucial restriction to be imposed is the following:

(H4) There exists a nondecreasing scalar function $\gamma(\|\xi\|) \geqq 0$, bounded for all bounded ξ, such that

$$\|G(x, \xi)\| \leqq \gamma(\|\xi\|)\|x\| \quad \text{for all } x, \xi. \tag{4.12}$$

This assumption is much less restrictive than the linear growth condition of [18]. It includes, for example, the product nonlinearities such as $G(x, \xi)\xi = \xi^2 x$.

PROPOSITION 5. *If $x = 0$ is the* GES *equilibrium of $\dot{x} = f(x, 0)$ and* (H1) *and* (H4) *hold, then the equilibrium $(x, \xi) = (0, 0)$ of the composite feedback system* (4.1) *is* GES *for every linear feedback $u = K\xi$ such that* $\operatorname{Re} \lambda(A_K) < 0$.

Proof. In view of the GES assumption, the Lyapunov function $V(x)$ defined in (H2) has the following additional properties:

$$\alpha_1\|x\|^2 \leqq V(x) \leqq \alpha_2\|x\|^2, \qquad \|\nabla V(x)\| \leqq \alpha_3\|x\|, \tag{4.13}$$

$$\dot{V} \leqq -\alpha_0 V, \tag{4.14}$$

where $\dot{V}$ is the derivative of V for $\dot{x} = f(x, 0)$ and $\alpha_0, \cdots, \alpha_3$ are some positive constants. Taking the derivative of V for (4.11b) we obtain

$$\dot{V}(x, t) = \nabla V(x) f(x, 0) + \nabla V(x) G(x, \xi(t))\xi(t), \tag{4.15}$$

where any solution $\xi(t)$ of (4.11b) satisfies

$$\|\xi(t)\| \leqq k\, e^{-at}\|\xi(0)\|, \qquad k \geqq 1, \quad a > 0. \tag{4.16}$$

Taking into account (4.12), (4.13), (4.14), and (4.16) we obtain from (4.15)

$$\dot{V} \leqq -\alpha_0 V + \frac{\alpha_3 \gamma(k\|\xi(0)\|)}{\alpha_1}\|\xi(0)\|\, e^{-at} V. \tag{4.17}$$

From this inequality it follows that $V(x(t))$ is bounded by

$$V(x(t)) \leqq k_0(\xi(0))\, e^{-\alpha_0 t} V(x(0)), \tag{4.18}$$

where $k_0(\xi(0)) = \exp\{(\alpha_3 k/\alpha_0 a)\gamma(k\|\xi(0)\|)\|\xi(0)\|\}$. This completes the proof of global exponential stability of (4.11). □

5. Conclusions. The two types of structure constraints imposed by the coordinate-free stability condition of Theorem 2 are, first, the interconnection structure constraint and, second, the linear stable right invertibility constraint. To examine the first constraint, consider a decomposition $f(x, \zeta) = f_0(x, \xi) + R(x, \xi)$ that is more general than (3.3). A simple extension of the assumption (H2) is to require for $\dot{x} = f_0(x, \xi)$ that the asymptotic stability property, guaranteed by $V(x)$, be uniform in ξ. Much more fundamental is the question of whether an assumption about the interconnection $R(x, \xi)$, less restrictive than (H3), can be made. Once a linear subsystem output $y = C\xi = \xi_1$ is chosen, the assumption (H3) disallows $R(x, \xi)$ to depend on linear variables other than ξ_1 and the zero dynamics ξ_0 induced by the output ξ_1. For linear systems with relative degree two and higher, this restriction is a challenging research topic. If, as our Example 3 suggests, the interconnection condition (H3) is in some cases necessary, then the challenge is to delineate such cases, and to search for less restrictive conditions for other classes of systems. In any event, the study of delicate interconnection properties, initiated in [11] and in this paper, is a promising direction for future research.

The second condition, which restricts the linear subsystem to be right invertible and weakly minimum phase, cannot be relaxed, without imposing some form of growth restriction on the nonlinear subsystem, as shown in Theorem 3 and Proposition 5. A direction in which the right invertibility condition can be generalized is to consider that both subsystems in the cascade are nonlinear and the first one is right invertible and globally minimum phase. The results of this paper combined with several nonlinear invertibility results starting with [8], justify the conjecture that a nonlinear analogue of Theorem 2 exists, at least for the minimum phase case.

Acknowledgment. Discussions with Alan Laub of the University of California at Santa Barbara have contributed to the final form of the Proposition 2.

REFERENCES

[1] A. ALBERT, *Conditions for positive and nonnegative definiteness in terms of pseudoinverse*, SIAM J. Appl. Math., 17 (1969), pp. 433-440.

[2] C. I. BYRNES AND A. ISIDORI, *A frequency domain philosophy for nonlinear systems, with application to stabilization and to adaptive control*, in Proc. 23rd Annual IEEE Conference on Decision and Control, IEEE Computer Society, Washington, D.C., 1984, pp. 1569-1573.

[3] ———, *Global feedback stabilization of nonlinear minimum phase systems*, in Proc. 24th Annual IEEE Conference on Decision and Control, IEEE Computer Society, Washington, D.C., 1985, pp. 1031-1037.

[4] ———, *Local stabilization of minimum phase nonlinear systems*, Systems Control Lett., 10 (1988), pp. 9-17.

[5] ———, *New results and examples in nonlinear feedback stabilization*, Systems Control Lett., 12 (1989), pp. 437-442.

[6] E. G. GILBERT, *The decoupling of multivariable systems by state feedback*, SIAM J. Control, 7 (1969), pp. 50-63.

[7] W. HAHN, *Stability of Motion*, Springer-Verlag, New York, 1967.

[8] R. M. HIRSCHORN, *Invertibility of multivariable nonlinear control systems*, IEEE Trans. Automat. Control, 24 (1979), pp. 865-885.

[9] A. ISIDORI AND C. H. MOOG, *On the nonlinear equivalent of the notion of transmission zeros*, in Modeling and Adaptive Control, C. I. Byrnes and A. Kurszanski, eds., Lecture Notes in Control and Information Sciences, vol. 105, Springer-Verlag, Berlin, New York, 1986.

[10] D. E. KODITSCHEK, *Adaptive techniques for mechanical systems*, in Proc. 5th Yale Workshop on Adaptive Systems, Yale University, New Haven, CT, 1987, pp. 259-265.

[11] P. V. KOKOTOVIC AND H. J. SUSSMANN, *A positive Real Condition for global stabilization of nonlinear systems*, Systems Control Lett., 13 (1989), pp. 125-133.

[12] R. MARINO, *High-gain feedback in non-linear control systems*, Internat. J. Control, 42 (1985), pp. 1369-1385.

[13] A. N. MICHEL AND R. K. MILLER, *Qualitative analysis of Large Scale Dynamical Systems*, Mathematics In Science and Engineering, Vol. 134, Academic Press, New York, 1977.

[14] A. S. MORSE, *Structural invariants of linear multivariable systems*, SIAM J. Control, 11 (1973), pp. 446-465.

[15] K. S. NARENDRA AND P. KUDVA, *Stable adaptive schemes for system identification and control—parts* I and II, IEEE Trans. Systems Man Cybernetics, 4 (1974), pp. 542-560.

[16] P. C. PARKS, *Lyapunov redesign of model reference adaptive systems*, IEEE Trans. Automat. Control, 11 (1966), pp. 362-369.

[17] P. SANNUTI, *Direct singular perturbation analysis of high-gain and cheap control problems*, Automatica, 19 (1983), pp. 41-51.

[18] S. SASTRY AND A. ISIDORI, *Adaptive control of linearizable systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 1123-1131.

[19] D. D. SILJAK, *Large-Scale Dynamic Systems, Stability and Structure*, North-Holland, Amsterdam, 1978.

[20] L. M. SILVERMAN, *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, 14 (1969), pp. 270-276.

[21] P. K. SINHA, *Dynamic compensation for state feedback decoupling of multivariable systems*, Internat. J. Control, 24 (1976), pp. 673-684.

[22] E. D. SONTAG, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435-443.

[23] P. SANNUTI AND A. SABERI, *A special coordinate basis of multivariable linear systems—finite and infinite zero structure, squaring down and decoupling*, Internat. J. Control, 45 (1987), pp. 1655-1704.

[24] A. SABERI AND P. SANNUTI, *Squaring down by static and dynamic compensators*, IEEE Trans. Automat. Control, 33 (1988), pp. 358-365.

[25] J. TSINIAS, *Sufficient Lyapunov like conditions for stabilization*, Math. Control Signals Syst., 2 (1989), pp. 343-357.

[26] M. VIDYASAGAR, *Decomposition techniques for large-scale systems with nonadditive interactions: stability and stabilizability*, IEEE Trans. Automat. Control, 25 (1980), pp. 773-779.

[27] W. A. WOLOVICH, *Linear Multivariable Systems*, Springer-Verlag, Berlin, 1974.